

Numeri sulla pietra: epigrafia quantitativa, statistica delle iscrizioni e big data dell'antichità

L'epigrafia nell'era dei dati massivi

L'epigrafia, disciplina tradizionalmente consacrata allo studio filologico e storico delle iscrizioni antiche, ha conosciuto negli ultimi decenni una trasformazione radicale che ne ha ridefinito profondamente i metodi, gli obiettivi e le possibilità euristiche. Quella che per secoli è stata un'arte interpretativa fondata sulla lettura paziente di singoli testi incisi sulla pietra, sul bronzo o su altri materiali durevoli, si è progressivamente evoluta in una scienza quantitativa capace di interrogare simultaneamente decine o centinaia di migliaia di documenti epigrafici, estraendo patterns statistici, tendenze temporali, distribuzioni geografiche e reti sociali che sfuggono completamente all'osservazione individuale dei testi.

Questa rivoluzione metodologica non rappresenta un abbandono delle competenze filologiche tradizionali, ma piuttosto un loro potenziamento attraverso l'integrazione di strumenti computazionali, statistici e visuali che permettono di porre domande completamente nuove al materiale epigrafico. Se l'epigrafo classico si interrogava sul significato di una singola abbreviazione o sulla carriera di un personaggio attestato in un'iscrizione, l'epigrafo quantitativo può chiedersi quale sia stata l'evoluzione demografica di un'intera provincia nell'arco di tre secoli, quali fossero le reti di patronato che strutturavano la società municipale romana, come si diffondevano geograficamente i culti orientali attraverso l'analisi delle dediche votive, o quale fosse la distribuzione della ricchezza nelle diverse regioni dell'impero basandosi sui costi dei monumenti funerari.

Il passaggio dall'epigrafia tradizionale a quella quantitativa è stato reso possibile da tre sviluppi convergenti verificatisi principalmente tra gli anni Ottanta del Novecento e i primi due decenni del Ventunesimo secolo. In primo luogo, la digitalizzazione sistematica dei grandi corpora epigrafici, un'impresa titanica che ha trasformato le migliaia di volumi stampati del *Corpus Inscriptionum Latinarum* e delle *Inscriptiones Graecae* in database interrogabili elettronicamente. In secondo luogo, lo sviluppo di tecniche statistiche e computazionali sempre più sofisticate, molte delle quali derivate dalla linguistica computazionale, dalla network science, dalla statistica spaziale e dal machine learning. Infine, la crescente disponibilità di potenza di calcolo e di software specializzati che rendono queste tecniche accessibili anche a ricercatori non necessariamente formati come statistici o informatici.

I corpus epigrafici come datasets strutturati: dalla pietra al database

La trasformazione dell'epigrafia in scienza quantitativa inizia necessariamente dalla costruzione di datasets strutturati che contengano non soltanto il testo delle iscrizioni, ma anche una ricca serie di metadati che ne descrivano le caratteristiche materiali, cronologiche, geografiche e contestuali. I tre principali progetti di digitalizzazione epigrafica che hanno reso possibile questa trasformazione sono il *Corpus Inscriptionum Latinarum* nella sua incarnazione digitale, le *Inscriptiones Graecae* e soprattutto la *Epigraphic Database Heidelberg* (EDH), che rappresenta attualmente il più avanzato esempio di database epigrafico relazionale.

Il *Corpus Inscriptionum Latinarum*, monumentale impresa filologica avviata da Theodor Mommsen nel 1853 e tuttora in corso di pubblicazione, ha raccolto e pubblicato in forma critica oltre quattrocentomila iscrizioni latine provenienti da tutto l'impero romano. La sua digitalizzazione, iniziata negli anni Novanta, ha trasformato questa enorme massa di testi da una serie di volumi cartacei consultabili solo nelle maggiori biblioteche di ricerca in un database interrogabile online. Ogni iscrizione nel CIL digitale è accompagnata da metadati che includono la sua localizzazione geografica originaria (quando nota), la datazione proposta dagli editori (spesso espressa come intervallo temporale), il tipo di supporto materiale (stele, ara, lastra, elemento architettonico, sarcofago, cippo, ecc.), la categoria tipologica (epigrafe funeraria, onoraria, votiva, editto imperiale, milestone, ecc.), e naturalmente il testo stesso, sia nella forma abbreviata originale che in quella sciolta secondo le convenzioni epigrafiche.

La *Epigraphic Database Heidelberg*, sviluppata presso l'Università di Heidelberg e diretta per molti anni da Géza Alföldy, rappresenta un passo ulteriore nella strutturazione dei dati epigrafici. Contenendo attualmente oltre ottantacinquemila iscrizioni latine con documentazione fotografica e metadati estremamente dettagliati, l'EDH adotta un modello di database relazionale che permette interrogazioni complesse basate su molteplici criteri simultanei. È possibile, ad esempio, estrarre tutte le iscrizioni funerarie datate tra il 100 e il 200 d.C., provenienti dalla provincia della Gallia Narbonensis, che menzionino veterani della legio X Gemina, ordinate per località di rinvenimento. Questo tipo di interrogazione, che richiederebbe mesi di lavoro manuale consultando i volumi cartacei, viene eseguito dal database in pochi secondi.

La struttura relazionale dell'EDH prevede diverse tabelle interconnesse: una tabella principale delle iscrizioni, collegata a tabelle di persone menzionate (con i loro status giuridici, titolature, cariche), luoghi (con coordinate geografiche), datazioni (con espressione di incertezza), materiali e supporti, editori e bibliografia. Questa architettura permette non soltanto il retrieval delle informazioni, ma anche analisi quantitative sofisticate. La presenza di coordinate geografiche georeferenziate per la maggior parte delle iscrizioni consente analisi spaziali; la standardizzazione dei nomi di persona facilita l'onomastica quantitativa; la categorizzazione tipologica permette analisi comparative tra diverse classi di documenti.

Un aspetto cruciale della strutturazione dei dati epigrafici riguarda il trattamento dell'incertezza e della lacunosità. A differenza di molti altri tipi di dati storici, le iscrizioni antiche sono frequentemente frammentarie, erose dal tempo, deliberatamente danneggiate, o semplicemente incomplete. I sistemi di markup epigrafico, standardizzati attraverso le convenzioni dell'*Epidoc* (Epigraphic Documents in TEI XML), permettono di codificare formalmente questi diversi gradi di certezza e completezza. Le lettere incerte vengono

marcate, le integrazioni proposte vengono distinte dal testo sicuramente leggibile, le lacune vengono quantificate quando possibile. Questa codifica formale dell'incertezza è fondamentale per l'analisi quantitativa: permette infatti di assegnare pesi probabilistici alle diverse interpretazioni e di propagare l'incertezza attraverso le catene inferenziali.

La qualità dei metadati costituisce il vero valore aggiunto dei database epigrafici rispetto alle semplici raccolte di testi. Una datazione proposta come "metà del secondo secolo d.C." deve essere tradotta in un intervallo numerico (diciamo 130-170 d.C.) con un'indicazione del grado di certezza. Una provenienza descritta come "trovata nei pressi di Aquileia, probabilmente proveniente dal foro" deve essere georeferenziata con coordinate che riflettano questa incertezza spaziale. Un supporto descritto come "stele funeraria in calcare locale, altezza residua 87 cm" deve essere categorizzato secondo schemi standardizzati che permettano confronti quantitativi. La costruzione di questi metadati richiede un lavoro filologico ed archeologico estremamente accurato, che rappresenta forse il contributo più importante dell'epigrafia tradizionale all'epigrafia quantitativa.

Distribuzione spaziale e analisi geografica: cartografie del passato

Una delle applicazioni più immediate e visivamente efficaci dell'epigrafia quantitativa riguarda l'analisi della distribuzione geografica delle iscrizioni e l'inferenza di fenomeni storici, culturali ed economici da questi patterns spaziali. La disponibilità di coordinate geografiche per centinaia di migliaia di iscrizioni permette di costruire mappe di densità, di identificare clusters significativi, di tracciare gradienti di diffusione culturale, e di testare ipotesi sulla struttura territoriale dell'impero romano.

La tecnica statistica della kernel density estimation (KDE) si rivela particolarmente utile per visualizzare la distribuzione delle iscrizioni nello spazio. A differenza di una semplice mappa a punti, che può risultare illeggibile quando migliaia di iscrizioni sono concentrate in aree urbane, la KDE produce una superficie di densità continua che evidenzia le regioni di maggiore concentrazione epigrafica. Il metodo funziona assegnando a ogni iscrizione una funzione kernel (tipicamente gaussiana) centrata sulle sue coordinate, e sommando poi tutti questi kernel per produrre una stima della densità in ogni punto dello spazio. La larghezza del kernel (bandwidth) rappresenta un parametro cruciale: un bandwidth troppo piccolo produce una superficie estremamente irregolare che riflette semplicemente la posizione esatta di ogni iscrizione, mentre un bandwidth troppo ampio produce una superficie eccessivamente smoothed che oscura variazioni significative.

Applicando la KDE alle iscrizioni latine dell'impero romano, emergono patterns attesi ma anche sorprese significative. Come prevedibile, le maggiori concentrazioni si trovano in corrispondenza delle grandi città: Roma naturalmente domina con decine di migliaia di iscrizioni, seguita da centri come Ostia, Pozzuoli, Capua in Italia, Lione e Nîmes in Gallia, Tarragona in Spagna, Lambèse e Thamugadi in Africa, Efeso e Afrodizia in Asia Minore. Tuttavia, l'analisi quantitativa rivela anche concentrazioni sorprendenti in aree meno urbanizzate ma culturalmente significative: la Dalmazia, ad esempio, presenta una densità epigrafica straordinariamente alta rispetto al suo livello di urbanizzazione, riflettendo

probabilmente un'intensa penetrazione culturale romana tra popolazioni illiriche che adottarono con entusiasmo la pratica della commemorazione epigrafica.

La distribuzione geografica delle iscrizioni non riflette semplicemente la densità di popolazione antica, ma piuttosto una complessa interazione tra fattori demografici, culturali ed economici. Regioni densamente popolate ma culturalmente meno romanizzate (come l'Egitto rurale) hanno prodotto relativamente poche iscrizioni latine, mentre regioni periferiche ma fortemente militarizzate (come la Britannia settentrionale o la frontiera renana) mostrano concentrazioni significative associate alle guarnigioni legionarie. L'Africa proconsolare presenta un caso particolarmente interessante: la densità epigrafica è straordinariamente alta non solo nelle grandi città costiere ma anche nei centri agrari dell'interno, riflettendo la prosperità economica della regione durante il secondo e terzo secolo e l'adozione entusiastica della cultura urbana romana da parte delle élites municipali.

L'analisi spaziale permette anche di tracciare fenomeni di diffusione culturale attraverso lo spazio e il tempo. Considerando ad esempio le iscrizioni che menzionano specifici culti orientali (Iside, Mitra, Cibele, Giove Dolicheno), è possibile mappare la loro distribuzione geografica e, incrociando i dati spaziali con quelli cronologici, ricostruire i percorsi e i tempi della loro diffusione nell'impero. Il culto mitraico, ad esempio, mostra una distribuzione fortemente correlata con le rotte commerciali fluviali e con le zone di concentrazione militare, suggerendo che i mercanti e i soldati furono i principali vettori della sua diffusione. Le iscrizioni di Iside, invece, si concentrano inizialmente nei porti mediterranei e solo successivamente penetrano nell'interno, tracciando chiaramente le rotte del commercio marittimo.

Tecniche più sofisticate di spatial statistics permettono di andare oltre la semplice visualizzazione. L'analisi dei nearest neighbor distances può testare statisticamente se la distribuzione delle iscrizioni è casuale, aggregata o dispersa. La funzione K di Ripley permette di identificare scale spaziali caratteristiche di clustering. I modelli di point process spaziali possono incorporare covariate ambientali (distanza da strade romane, prossimità a centri urbani, altitudine) per modellare i fattori che influenzano la densità epigrafica. Questi approcci hanno rivelato, ad esempio, che le iscrizioni miliari (pietre miliari) non sono distribuite uniformemente lungo le strade romane, ma tendono a concentrarsi in prossimità di centri abitati e di confini amministrativi, riflettendo il loro duplice ruolo di indicatori stradali e di marcatori simbolici di sovranità imperiale.

Un'applicazione particolarmente innovativa dell'analisi spaziale riguarda lo studio delle reti commerciali attraverso la provenienza dei materiali lapidei. Molte iscrizioni romane, specialmente quelle di carattere monumentale, erano incise su marmi pregiati importati da cave specifiche: marmo di Carrara (Luna), marmo pentelico da Atene, marmo proconnesio dalla Propontide, porfido rosso dall'Egitto, granito dall'Assuan. L'analisi petrografica permette di identificare la provenienza geologica di questi materiali, e la georeferenziazione delle iscrizioni permette quindi di tracciare le rotte commerciali che collegavano le cave ai luoghi di destinazione. Emerge così una vera e propria network of stone, una rete di flussi commerciali che collega miniere, porti, centri di lavorazione e città destinatarie. L'analisi di questa rete rivela la straordinaria integrazione economica dell'impero romano: marmo greco veniva esportato in Britannia, granito egiziano raggiungeva la Spagna, calcare nordafricano veniva lavorato in Italia.

Tendenze temporali: l'epigrafia come termometro della storia

Se l'analisi spaziale permette di cartografare la romanità, l'analisi temporale della produzione epigrafica fornisce una sorta di termometro della vitalità culturale, economica e demografica delle diverse regioni dell'impero attraverso i secoli. La frequenza con cui le iscrizioni venivano prodotte in un dato periodo non è costante, ma fluttua significativamente nel tempo, riflettendo cambiamenti nella struttura sociale, nella disponibilità di risorse economiche, nelle pratiche commemorative, e più fondamentalmente nella stessa popolazione che poteva permettersi di commissionare monumenti epigrafici.

Costruire una serie temporale della produzione epigrafica richiede innanzitutto di confrontarsi con il problema fondamentale della datazione. A differenza dei documenti scritti su papiro o pergamena, che spesso recano date consolari esplicite, la maggior parte delle iscrizioni non contiene elementi cronologici precisi. Gli epigrafi devono quindi ricorrere a metodi indiretti: la paleografia (lo studio della forma delle lettere, che evolve nel tempo), la prosopografia (l'identificazione di personaggi noti da altre fonti), la tipologia dei monumenti, la formulazione del testo, il contesto archeologico. Queste tecniche permettono generalmente di datare un'iscrizione con una precisione che varia da pochi decenni a un secolo intero, ma raramente a singoli anni.

Questa incertezza cronologica pone sfide significative per l'analisi quantitativa. Una soluzione consiste nell'aggregare le iscrizioni in bins temporali sufficientemente ampi (per esempio mezzi secoli o secoli interi), sacrificando risoluzione temporale in cambio di robustezza statistica. Un approccio più sofisticato utilizza invece metodi bayesiani per rappresentare l'incertezza cronologica come distribuzioni di probabilità: se un'iscrizione è datata "probabilmente prima metà del secondo secolo", questa informazione viene codificata come una distribuzione di probabilità concentrata tra il 100 e il 150 d.C. ma con code che si estendono prima e dopo. L'aggregazione di molte iscrizioni con incertezze cronologiche può poi produrre stime robuste delle tendenze temporali.

Quando si costruisce una serie temporale $f(t)$ della frequenza epigrafica, emerge un pattern sorprendentemente consistente attraverso diverse regioni dell'impero, con variazioni locali significative ma sovrapposte a una tendenza generale comune. La produzione epigrafica inizia a crescere significativamente nel tardo primo secolo a.C., in coincidenza con l'età augustea e con la stabilizzazione politica seguita alle guerre civili. Questa crescita continua attraverso tutto il primo secolo d.C., accelera ulteriormente nel secondo secolo, e raggiunge il suo picco assoluto nell'età antonina, approssimativamente tra il 140 e il 180 d.C. Questo periodo, che coincide con il regno di Antonino Pio e la prima parte del regno di Marco Aurelio, rappresenta davvero l'apogeo della cultura epigrafica romana.

Il picco antonino della produzione epigrafica non è uniforme geograficamente: è particolarmente pronunciato in Africa proconsolare e in alcune province orientali, meno marcato in Italia dove la produzione era già elevata nel primo secolo, moderato nelle province galliche e iberiche. Questa variazione geografica riflette probabilmente differenze nei ritmi di sviluppo economico e di integrazione culturale: l'Africa raggiunge la sua massima

prosperità nel secondo secolo, mentre l'Italia sperimenta già nel primo secolo quella ricchezza che in Africa maturerà solo successivamente.

Dopo il picco antonino, la produzione epigrafica inizia un declino graduale ma inesorabile che continua attraverso tutto il terzo secolo e accelera drammaticamente nel quarto. Le cause di questo declino rappresentano uno dei problemi più dibattuti dell'epigrafia quantitativa. Alcuni studiosi hanno interpretato la diminuzione delle iscrizioni come riflesso di un declino economico e demografico generale (l'"anarchia militare" del terzo secolo), ma questa interpretazione è stata contestata sulla base del fatto che l'economia tardoantica non era necessariamente in declino rispetto all'alto impero. Una spiegazione alternativa enfatizza i cambiamenti culturali: la graduale cristianizzazione dell'impero portò a diverse pratiche commemorative (l'iscrizione funeraria cristiana è tipicamente molto più semplice e standardizzata di quella pagana), e più in generale la cultura tardoantica sembra aver investito meno risorse nella auto-rappresentazione epigrafica monumentale caratteristica dell'alto impero.

L'analisi quantitativa permette di testare queste diverse ipotesi esaminando quali categorie di iscrizioni declinano più rapidamente. Se il declino fosse puramente economico, dovremmo aspettarci che tutte le categorie diminuiscano proporzionalmente. Se invece è principalmente culturale, dovremmo osservare declini differenziali. I dati mostrano effettivamente che le iscrizioni funerarie private declinano più rapidamente delle iscrizioni pubbliche e monumentali, suggerendo un cambiamento nelle pratiche commemorative private piuttosto che un collasso generale della capacità economica. Le iscrizioni imperiali e i milestones continuano infatti ad essere prodotti a ritmi elevati fino al quarto secolo avanzato, mentre le stele funerarie private quasi scompaiono già nel terzo secolo.

Un'applicazione particolarmente sofisticata dell'analisi temporale riguarda la stima della popolazione antica attraverso i modelli di survival analysis applicati ai dati epigrafici. Le iscrizioni funerarie romane spesso riportano l'età al momento della morte, fornendo così un campione (certamente non rappresentativo) della mortalità antica. Applicando tecniche di survival analysis e correggendo per i bias sistematici (le iscrizioni tendono a sovra-rappresentare morti premature che erano considerate particolarmente tragiche, e sotto-rappresentare bambini molto piccoli e schiavi), è possibile stimare curve di sopravvivenza e aspettative di vita. Ancora più interessante, confrontando le curve di mortalità estratte da iscrizioni datate a diversi periodi, è possibile tracciare cambiamenti temporali nelle condizioni demografiche.

Text mining e Named Entity Recognition: estrarre informazioni strutturate dal testo

Oltre ai metadati esplicitamente codificati nei database epigrafici, un'enorme quantità di informazione strutturata è implicitamente contenuta nel testo stesso delle iscrizioni e può essere estratta attraverso tecniche di text mining e natural language processing. Le iscrizioni romane, specialmente quelle più formali e monumentali, seguono formulari relativamente standardizzati che facilitano l'estrazione automatica di informazioni. Una tipica iscrizione funeraria romana potrebbe seguire questo schema: nome del defunto con filiazione e tribù, cariche ricoperte in ordine cronologico (cursus honorum), età al momento della morte, nomi

dei dedicanti, formula di dedica. Un'iscrizione onoraria seguirà un pattern simile ma con l'aggiunta della motivazione dell'onore e dell'indicazione dell'organo che lo ha conferito.

La Named Entity Recognition (NER) è una tecnica di NLP che mira ad identificare e classificare automaticamente entità nominate in un testo: nomi di persone, luoghi, organizzazioni, date, titoli e cariche. Applicata ai testi epigrafici, la NER può estrarre automaticamente migliaia di informazioni strutturate che sarebbero altrimenti disperse nel testo non strutturato. Addestrare un sistema di NER per l'epigrafia latina richiede la creazione di un corpus annotato dove le diverse categorie di entità sono state marcate manualmente da esperti epigrafisti. Questo gold standard viene poi utilizzato per addestrare modelli di machine learning (tipicamente basati su Conditional Random Fields o, più recentemente, su reti neurali ricorrenti come LSTM o transformer) che imparano a riconoscere le caratteristiche linguistiche e contestuali che segnalano la presenza di una determinata entità.

Le sfide specifiche della NER applicata all'epigrafia latina includono l'alto grado di abbreviazione (che richiede un sistema di espansione automatica delle abbreviazioni), la variabilità ortografica (il latino epigrafico presenta numerose varianti grafiche e "errori" che riflettono pronunce locali), la lacunosità dei testi (che richiede di gestire testo parzialmente mancante), e la necessità di disambiguare nomi identici (molti individui condividono gli stessi tria nomina). Nonostante queste sfide, i sistemi di NER applicati all'epigrafia hanno raggiunto livelli di accuratezza superiori al novanta per cento per categorie ben definite come i nomi personali e le cariche magistratuali.

L'estrazione automatica di nomi personali dalle iscrizioni alimenta direttamente l'onomastica quantitativa, uno dei campi più fecondi dell'epigrafia computazionale. Il sistema onomastico romano, con la sua struttura tripartita di praenomen, nomen e cognomen (più eventuali agnomina), codifica informazioni sulla famiglia, sulla tribù, sull'origine geografica e talvolta sullo status sociale dell'individuo. Analizzando la distribuzione di migliaia di nomi estratti automaticamente dalle iscrizioni, è possibile tracciare fenomeni demografici, migratori e culturali. La frequenza relativa dei diversi praenomina, ad esempio, mostra variazioni geografiche e temporali significative: il praenomen Marcus è particolarmente comune in alcune aree mentre Titus domina in altre, riflettendo probabilmente tradizioni familiari regionali.

Lo studio dei cognomina è ancora più rivelatore. Molti cognomina hanno origini etniche chiare: cognomina greci (terminanti in -as, -os, -is) sono comuni tra liberti di origine orientale; cognomina celtici sono prevalenti in Gallia; cognomina punici persistono in Africa anche in epoca imperiale. La distribuzione geografica di questi cognomina permette di tracciare patterns migratori e di integrazione culturale. Un fenomeno particolarmente interessante è la diffusione dei cognomina imperiali: dopo che un imperatore adottava un certo cognomen, questo tendeva a diventare molto più popolare nella popolazione generale, riflettendo sia l'adulazione politica sia genuine mode culturali.

L'estrazione automatica di cariche magistratuali e militari dalle iscrizioni permette di ricostruire careers patterns e di analizzare quantitativamente il cursus honorum romano. Le iscrizioni onorarie e funerarie di senatori e cavalieri spesso elencano dettagliatamente tutte le cariche ricoperte, fornendo una documentazione incomparabile sulla struttura

amministrativa e militare dell'impero. Aggregando queste informazioni da migliaia di iscrizioni, è possibile ricostruire le sequenze tipiche di cariche, identificare percorsi di carriera alternativi, quantificare le età tipiche di accesso alle diverse magistrature, e tracciare cambiamenti nel tempo nella struttura burocratica imperiale. Ad esempio, l'analisi quantitativa ha rivelato che il cursus senatorio subì trasformazioni significative tra il primo e il terzo secolo, con una progressiva militarizzazione e un aumento delle carriere "equestri" a scapito di quelle strettamente senatorie.

L'estrazione di informazioni geografiche dal testo delle iscrizioni complementa i metadati di provenienza. Molte iscrizioni menzionano luoghi: la patria di origine di un defunto migrante, la provincia dove un soldato ha servito, le città che hanno beneficiato della generosità di un evergete. Estraendo e geocodificando automaticamente questi riferimenti geografici, è possibile costruire networks di connessioni tra luoghi che riflettono mobilità, commercio e relazioni politiche. Un veterano della legio III Augusta morto e sepolto a Roma ma la cui iscrizione menziona la sua origine da Thugga in Africa crea un link tra questi due luoghi in una rete di mobilità militare.

Social network analysis: la prosopografia come grafo

Una delle applicazioni più innovative e concettualmente ricche dell'epigrafia quantitativa consiste nel trasformare le informazioni prosopografiche contenute nelle iscrizioni in reti sociali formali, analizzabili attraverso le tecniche della network science. Le iscrizioni romane documentano numerosi tipi di relazioni sociali: legami familiari (matrimoni, filiazioni, adozioni), relazioni di patronato (tra patroni e liberti o tra patroni politici e clienti), associazioni professionali e religiose (collegia), collaborazioni politiche (co-magistrature), e legami militari (soldati della stessa unità, ufficiali e subordinati). Ognuna di queste relazioni può essere rappresentata formalmente come un grafo, dove i nodi rappresentano individui e gli archi rappresentano relazioni tra di loro.

La costruzione di un grafo prosopografico a partire dalle iscrizioni richiede diverse operazioni preliminari. Innanzitutto, è necessario disambiguare i nomi: lo stesso tria nomina può riferirsi a individui diversi, oppure lo stesso individuo può essere menzionato con varianti del suo nome in iscrizioni diverse. Questa disambiguazione richiede di combinare criteri cronologici (due individui con lo stesso nome che vissero in epoche diverse sono certamente persone distinte), geografici (è improbabile che la stessa persona appaia contemporaneamente in iscrizioni da province remote), prosopografici (la menzione delle stesse cariche o degli stessi parenti aiuta a identificare lo stesso individuo), e statistici (modelli probabilistici possono assegnare likelihood che due menzioni si riferiscano alla stessa persona).

Una volta costruito il grafo, le tecniche di network analysis permettono di estrarre informazioni strutturali che sarebbero impossibili da cogliere attraverso l'esame individuale delle iscrizioni. Le misure di centralità identificano gli individui più importanti nella rete sociale. La betweenness centrality, che misura quante volte un nodo si trova su cammini minimi tra coppie di altri nodi, identifica individui che fungevano da bridge tra diversi gruppi sociali. La eigenvector centrality, che assegna importanza maggiore a nodi collegati ad altri nodi importanti, identifica individui centrali nell'élite. Il degree centrality semplicemente conta il numero di connessioni, identificando gli individui più socialmente attivi.

Applicando queste misure alla rete prosopografica dell'ordine senatorio romano ricostruita dalle iscrizioni, emergono risultati che illuminano la struttura del potere imperiale. Non sorprendentemente, i membri della famiglia imperiale mostrano altissimi valori di centralità, ma ancora più interessanti sono gli individui non imperiali che raggiungono comunque alti livelli di centralità: questi sono tipicamente uomini politici influenti che ricoprirono multiple cariche, costruirono ampie reti di patronato, e contrattarono matrimoni strategici per i loro discendenti. La rete rivela anche l'esistenza di fazioni: gruppi densamente connessi internamente ma poco connessi con altri gruppi, che probabilmente riflettono alleanze politiche, regionalismi, o orientamenti ideologici.

L'analisi di community detection può identificare automaticamente questi gruppi. Algoritmi come il Louvain method o il label propagation cercano di partizionare la rete in comunità che massimizzano la densità di connessioni interne rispetto a quelle esterne. Applicati alla rete senatoria, questi algoritmi tendono a identificare comunità basate su criteri geografici (senatori originari dalla stessa provincia tendono a essere più connessi tra loro), familiari (grandi gentes formano comunità distinte), e talvolta cronologici (cambiamenti di regime politico possono portare a turnover nella composizione dell'élite, creando clusters temporali).

Un'applicazione particolarmente affascinante della network analysis prosopografica riguarda lo studio dei matrimoni senatoriali come rete di alleanze. Ogni matrimonio tra famiglie senatorie crea un link nel grafo, e l'analisi della struttura di questo grafo rivela strategie matrimoniali e equilibri di potere. Alcune famiglie adottano strategie di endogamia, sposandosi ripetutamente all'interno di un ristretto gruppo di famiglie alleate, creando componenti densamente connesse nella rete. Altre famiglie seguono strategie più diversificate, cercando alleanze con gruppi diversi e posizionandosi come bridges. L'analisi temporale di questa rete rivela come le strategie matrimoniali cambino in risposta a eventi politici: dopo guerre civili o cambi di dinastia, si osservano spesso ristrutturazioni significative della rete matrimoniale.

La rete di patronato tra patroni e liberti, ricostruibile dalle migliaia di iscrizioni che menzionano liberti e i loro patroni, rivela la struttura capillare del potere sociale ed economico romano. Alcune famiglie ricche controllavano vere e proprie reti clientelari di centinaia di liberti, che a loro volta potevano avere propri liberti, creando strutture gerarchiche a più livelli. L'analisi di queste reti rivela che il potere sociale non dipendeva solo dalla ricchezza o dal rango formale, ma anche dalla capacità di costruire e mantenere ampie reti di dipendenti e obbligati.

Le reti professionali ricostruibili dalle iscrizioni dei collegia (associazioni di mestiere, culturali, funerarie) mostrano patterns diversi. Queste reti tendono ad essere più egualitarie, con distribuzione dei gradi più uniformi e meno gerarchiche delle reti di patronato. Tuttavia, anche all'interno dei collegia emergono individui con ruoli di leadership che occupano posizioni di centralità: i patroni dei collegia, i quinquennales (presidenti quinquennali), i curatores. L'analisi rivela che questi leader spesso fungevano da bridges tra il mondo dei collegia e le élite municipali o imperiali, facilitando risorse e protezione politica in cambio di supporto elettorale o sociale.

Un aspetto metodologicamente sofisticato della social network analysis prosopografica riguarda la gestione dell'incompletezza dei dati. La rete che possiamo ricostruire dalle

iscrizioni è certamente solo una frazione della vera rete sociale antica: molte relazioni non venivano commemorate epigraficamente, molte iscrizioni sono andate perdute, e molte zone geografiche sono sotto-rappresentate nel record epigrafico. Metodi di network inference possono tentare di ricostruire links mancanti basandosi su patterns strutturali: se due individui condividono molti vicini comuni nella rete, è probabile che fossero essi stessi connessi anche se non abbiamo evidenza epigrafica diretta di questo link. Tecniche bayesiane permettono di quantificare l'incertezza in queste inferenze.

Le reti prosopografiche possono anche essere analizzate dinamicamente, tracciando la loro evoluzione nel tempo. Costruendo snapshots della rete a intervalli temporali regolari, è possibile osservare come la struttura sociale cambi: nuove famiglie emergono e vecchie declinano, la densità della rete può aumentare o diminuire, la distribuzione della centralità può diventare più o meno diseguale. Queste dinamiche riflettono processi storici di mobilità sociale, cambiamenti politici, e trasformazioni economiche. Ad esempio, l'analisi dinamica della rete senatoria mostra un aumento significativo della mobilità sociale tra il primo e il secondo secolo, con maggiori opportunità per homines novi di accedere alle alte magistrature, seguito da una cristallizzazione crescente nel terzo secolo.

Topic modeling: temi ricorrenti nelle iscrizioni

Oltre all'estrazione di entità nominate e relazioni strutturate, le tecniche di text mining permettono di identificare i temi ricorrenti nei testi epigrafici e di tracciare la loro distribuzione geografica e temporale. Il topic modeling, particolarmente nella sua implementazione come Latent Dirichlet Allocation (LDA), fornisce un framework formale per scoprire automaticamente temi latenti in grandi corpora di testi. Applicato alle iscrizioni romane, il topic modeling può rivelare patterns tematici che attraversano migliaia di testi, identificare categorie di iscrizioni basate sul loro contenuto piuttosto che sulla loro tipologia formale, e tracciare l'evoluzione dei discorsi epigrafici nel tempo e nello spazio.

Il modello LDA assume che ogni documento (nel nostro caso, ogni iscrizione) sia generato da una mistura di topics, e che ogni topic sia caratterizzato da una distribuzione di probabilità sulle parole del vocabolario. L'algoritmo inferisce simultaneamente quali sono i topics e quale è la composizione di topics per ogni documento, basandosi unicamente sulla co-occorrenza delle parole nei testi. Un topic potrebbe essere caratterizzato da alte probabilità per parole come "consul", "legatus", "praetor", "tribunus", identificandosi come il topic delle cariche magistratuali. Un altro topic potrebbe avere alte probabilità per "pia", "dulcis", "carissimus", "benemerens", identificandosi come il topic degli epitaffi affettuosi.

Applicando LDA a un corpus di decine di migliaia di iscrizioni latine con un numero di topics fissato a venti o trenta, emergono topics interpretabili che corrispondono a categorie concettuali riconoscibili. Un topic cattura il linguaggio delle dediche votive agli dei, con parole come "votum", "solvit", "libens", "merito", "deo", "deae". Un altro topic cattura il linguaggio militare, con termini come "miles", "legio", "veteranus", "emeritus", "stipendia". Un topic cattura le formule di datazione, un altro le descrizioni di opere pubbliche, un altro ancora il linguaggio dei liberti e del manumissione.

Ciò che rende il topic modeling particolarmente potente è che permette di identificare topics che attraversano le categorie tipologiche tradizionali. Ad esempio, un topic potrebbe

emergere che cattura un discorso di pietas familiare che appare tanto in iscrizioni funerarie quanto in iscrizioni votive quanto in testi onorari, rivelando un tema ideologico trasversale. Un altro topic potrebbe catturare un linguaggio di evergetismo e beneficenza pubblica che unifica testi apparentemente diversi.

Una volta identificati i topics, è possibile analizzare la loro distribuzione geografica e temporale. Alcuni topics mostrano concentrazioni geografiche marcate: il topic delle cariche duovirali e quattuorvirali è molto più prevalente in Italia e nelle province occidentali che in quelle orientali, riflettendo differenze nella struttura amministrativa municipale. Il topic delle tribù è quasi assente nelle province orientali dove il sistema tribale romano non fu mai pienamente implementato. Il topic dei culti orientali mostra patterns di diffusione che possono essere mappati geograficamente.

L'analisi temporale dei topics rivela l'evoluzione dei discorsi epigrafici attraverso i secoli. Il topic del linguaggio imperiale augusteo (con termini come "Augustus", "divus", "pater patriae") emerge ovviamente solo a partire dall'età augustea e persiste attraverso l'impero. Il topic del linguaggio cristiano (con termini come "fidelis", "in pace", "depositus") emerge solo nel terzo secolo e diventa dominante nel quarto. Il topic delle virtù militari tradizionali è più prevalente nel primo secolo e declina progressivamente, mentre il topic della burocrazia imperiale cresce costantemente dal primo al terzo secolo.

Un'applicazione particolarmente interessante del topic modeling riguarda l'identificazione automatica di iscrizioni falsificate o sospette. Le false iscrizioni, create nel Rinascimento o nell'epoca moderna per varie ragioni (nazionalismo locale, frode, pura fantasia), spesso presentano anomalie nel loro profilo tematico: possono combinare topics in modi atipici, o usare un linguaggio che, pur sembrando "latino epigrafico", non corrisponde ai patterns statistici del vero corpus epigrafico. Modelli addestrati su iscrizioni sicuramente autentiche possono assegnare anomaly scores a iscrizioni sospette, aiutando gli studiosi a identificare probabili falsi.

Il topic modeling può anche essere utilizzato in modo semi-supervisionato, dove i ricercatori definiscono alcuni topics a priori basandosi su conoscenze epigrafiche (definendo ad esempio un seed set di parole caratteristiche per "iscrizioni funerarie cristiane"), e il modello poi identifica automaticamente altri documenti appartenenti a questa categoria. Questo approccio combina l'expertise filologica con la potenza computazionale, permettendo di classificare rapidamente migliaia di iscrizioni che sarebbero troppo numerose per una classificazione manuale completa.

Stilometria e attribuzione: cronologia attraverso lo stile formulare

La stilometria, lo studio quantitativo dello stile testuale, trova applicazioni particolarmente interessanti nell'epigrafia latina per due scopi principali: la datazione di iscrizioni non datate attraverso l'analisi delle loro caratteristiche stilistiche, e l'identificazione di officine lapidarie o di tradizioni formulari locali attraverso il riconoscimento di signature stilistiche distintive.

Le iscrizioni romane, specialmente quelle più formali, seguono formulari relativamente standardizzati che tuttavia evolvono nel tempo. La stilometria permette di catturare quantitativamente queste evoluzioni e di utilizzarle per la datazione. L'approccio classico si basa sull'identificazione di features stilistiche che sono note per cambiare temporalmente: la frequenza di certe abbreviazioni, l'ordine delle componenti formulari, la presenza o assenza di certe formule, la lunghezza media delle iscrizioni, la complessità sintattica.

Ad esempio, la formula dedicatoria "Dis Manibus" nelle iscrizioni funerarie subisce un'evoluzione temporale ben documentata. Nella sua forma più arcaica appare come "D(is) M(anibus) s(acrum)", poi si abbrevia progressivamente a "D(is) M(anibus)", poi semplicemente a "D(eis) M(anibus)", e infine in epoca tarda spesso viene omessa completamente o sostituita da formule cristiane. La presenza e la forma esatta di questa formula forniscono quindi indizi cronologici. Tuttavia, mentre un epigrafo esperto può riconoscere questi patterns intuitivamente, l'approccio stilometrico quantifica formalmente queste tendenze e le combina con centinaia di altre features stilistiche per produrre datazioni probabilistiche.

I metodi di machine learning supervisionato sono particolarmente adatti a questo compito. Si costruisce un training set di iscrizioni con datazioni sicure (magari perché menzionano consoli noti, o imperatori, o sono state datate archeologicamente), si estraggono features stilistiche da queste iscrizioni, e si addestra un modello (ad esempio un random forest, un support vector machine, o una rete neurale) a predire la data basandosi sulle features. Il modello impara automaticamente quali combinazioni di features sono predittive di quali periodi. Una volta addestrato, il modello può essere applicato a iscrizioni di data incerta per produrre stime cronologiche.

Le features stilistiche utilizzate possono essere di vario tipo. Features lessicali includono la frequenza di parole specifiche o gruppi di parole. Features sintattiche catturano la struttura grammaticale delle frasi. Features formulari identificano la presenza o assenza di formule standard e il loro ordine. Features di lunghezza includono il numero totale di parole, la lunghezza media delle parole, la distribuzione delle lunghezze delle frasi. Features di abbreviazione quantificano il grado di abbreviazione del testo.

Un aspetto particolarmente interessante della stilometria epigrafica riguarda l'identificazione di variazioni geografiche nello stile. Diverse regioni dell'impero svilupparono tradizioni formulari locali. Le iscrizioni africane tendono ad essere più prolisse e elaborate di quelle italiane; le iscrizioni della Britannia mostrano peculiarità ortografiche che riflettono pronunce locali; le iscrizioni della Gallia Narbonensis presentano influenze del formulario greco orientale più di quelle della Gallia settentrionale. La stilometria può quantificare queste differenze e utilizzarle non solo per datare iscrizioni, ma anche per identificarne la probabile provenienza geografica quando questa è incerta.

L'attribuzione stilometrica può anche identificare l'operato di specifiche officine lapidarie. Gli scalpellini professionisti che incidevano le iscrizioni sviluppavano stili caratteristici, sia nella paleografia (la forma delle lettere) che nel formulario e nel layout. Analizzando iscrizioni di provenienza geografica molto circoscritta, è possibile identificare clusters stilistici che probabilmente rappresentano il lavoro di singole officine o addirittura di singoli artigiani. Questo tipo di analisi può rivelare reti commerciali locali e circuiti di committenza.

Paleografia computazionale: classificazione automatica delle forme letterali

La paleografia, lo studio delle forme delle lettere e della loro evoluzione nel tempo, è tradizionalmente una delle competenze fondamentali dell'epigrafo. Le lettere capitali romane evolvono in modi documentabili attraverso i secoli: le lettere più antiche tendono ad essere più regolari e geometriche, quelle imperiali mostrano maggiore elaborazione decorativa con serif pronunciati, quelle tardoantiche mostrano tendenze verso forme più corsive e irregolari. Queste evoluzioni paleografiche forniscono importanti criteri di datazione, ma il loro riconoscimento ha sempre richiesto l'occhio esperto dell'epigrafo addestrato.

La paleografia computazionale mira ad automatizzare almeno parzialmente questo processo attraverso tecniche di image processing e machine learning applicato alla visione artificiale. Il workflow tipico inizia con l'acquisizione di immagini digitali di alta qualità delle iscrizioni, prosegue con la segmentazione automatica delle singole lettere, l'estrazione di features geometriche e topologiche da ogni lettera, e infine la classificazione delle lettere secondo criteri cronologici o tipologici attraverso algoritmi di machine learning.

La segmentazione delle lettere da un'immagine di un'iscrizione non è un compito banale. Le lettere possono essere variamente spaziate, possono toccarsi o sovrapporsi, possono essere parzialmente erose o danneggiate, possono variare in dimensione all'interno della stessa iscrizione. Algoritmi di segmentazione basati su tecniche di computer vision come la rilevazione di contorni, la trasformata di Hough per identificare linee, e metodi di watershed per separare regioni connesse sono stati applicati con discreto successo. Approcci più recenti utilizzano reti neurali convoluzionali (CNN) addestrate specificamente per la segmentazione di testo epigrafico.

Una volta segmentate le singole lettere, il passo successivo consiste nell'estrarre features che catturino le loro caratteristiche geometriche rilevanti per la datazione. Queste features possono includere: il rapporto altezza/larghezza delle lettere, la presenza e la forma di serif (le grazie decorative alle estremità dei tratti), la proporzione delle parti delle lettere (ad esempio, nelle lettere "E" o "F", la lunghezza relativa dei bracci orizzontali), l'angolo dei tratti obliqui, la regolarità e simmetria, il grado di modulazione (variazione di spessore) dei tratti, la presenza di elementi decorativi o ornamentali.

Diverse tecniche di feature extraction sono state sperimentate. L'approccio più diretto consiste nel calcolare manualmente features geometriche predefinite. Un approccio più sofisticato utilizza descrittori di forma standard dalla computer vision come HOG (Histogram of Oriented Gradients), SIFT (Scale-Invariant Feature Transform), o shape contexts. Gli approcci più recenti utilizzano il deep learning, dove le features vengono apprese automaticamente da reti neurali convoluzionali addestrate end-to-end sul compito di classificazione paleografica.

La classificazione paleografica può essere formulata come un problema di classificazione supervisionata multiclasse, dove le classi rappresentano periodi cronologici o stili paleografici. Un training set viene costruito selezionando iscrizioni con datazioni sicure e annotando le loro caratteristiche paleografiche. Features vengono estratte dalle lettere in queste iscrizioni, e un classificatore (ad esempio Support Vector Machine, Random Forest, o

Neural Network) viene addestrato a predire la classe paleografica basandosi sulle features. L'accuratezza della classificazione dipende criticamente dalla qualità e dalla quantità dei dati di training, oltre che dalla rilevanza delle features estratte.

I risultati finora ottenuti dalla paleografia computazionale sono promettenti ma non ancora tali da sostituire l'expertise umana. I migliori sistemi raggiungono accuratèzze del settanta-ottanta per cento nella datazione di iscrizioni a livello di mezzo secolo, una performance significativamente inferiore a quella di un epigrafo esperto ma comunque utile come strumento di screening preliminare per grandi corpora. Le difficoltà principali derivano dalla variabilità intrinseca della paleografia epigrafica: anche all'interno dello stesso periodo cronologico, diversi lapicidi mostravano stili diversi; la qualità del materiale lapideo e le capacità tecniche dell'incisore influenzavano il risultato; e fattori conservativi culturali potevano portare a persistenza di stili arcaici in alcuni contesti.

Un'applicazione particolarmente interessante della paleografia computazionale riguarda non tanto la datazione quanto l'identificazione di mani diverse all'interno di grandi complessi epigrafici. Ad esempio, nei grandi progetti epigrafici imperiali come i fasti trionfali o i cataloghi di membri di collegi, diverse sezioni potrebbero essere state incise da artigiani diversi in tempi diversi. L'analisi paleografica computazionale, attraverso clustering non supervisionato delle features paleografiche, può identificare automaticamente gruppi di iscrizioni con stili affini, suggerendo l'intervento delle stesse mani.

La fotogrammetria e la scansione 3D stanno aprendo nuove possibilità per la paleografia computazionale. Mentre le fotografie bidimensionali catturano solo l'aspetto delle lettere, le scansioni 3D catturano anche la loro profondità, l'angolo di incisione, la sezione del solco. Queste informazioni tridimensionali forniscono features aggiuntive potenzialmente molto informative sulle tecniche di incisione e sulle mani degli artigiani. Inoltre, le scansioni 3D permettono di "raddrizzare" virtualmente iscrizioni incise su superfici curve o irregolari, facilitando la lettura e l'analisi.

Survival analysis: la durata dei monumenti epigrafici

Un approccio quantitativo particolarmente originale all'epigrafia consiste nell'applicare tecniche di survival analysis, tradizionalmente utilizzate in biostatistica e ingegneria dell'affidabilità, allo studio della sopravvivenza fisica dei monumenti epigrafici attraverso i millenni. Ogni iscrizione che ci è pervenuta ha "sopravvissuto" a una lunga serie di potenziali eventi distruttivi: erosione naturale, terremoti, inondazioni, riutilizzo del materiale in nuove costruzioni, iconoclastia religiosa, depredazione di materiali pregiati, distruzione deliberata per ragioni politiche. Comprendere i fattori che influenzano la probabilità di sopravvivenza può illuminare non solo processi di conservazione archeologica, ma anche pratiche culturali antiche e post-antiche.

Il framework della survival analysis modella il tempo fino a un evento (in questo caso, la distruzione o la perdita dell'iscrizione) come una variabile aleatoria. La funzione di sopravvivenza $S(t)$ rappresenta la probabilità che un'iscrizione sopravviva oltre il tempo t . La funzione di hazard $h(t)$ rappresenta il tasso istantaneo di "fallimento" (distruzione) al tempo t , condizionato al fatto che l'iscrizione sia sopravvissuta fino a t . Queste funzioni possono

essere stimate empiricamente dai dati e modellate in funzione di covariate che caratterizzano le iscrizioni.

Per applicare la survival analysis alle iscrizioni antiche, dobbiamo innanzitutto definire cosa significa "sopravvivenza". Un'iscrizione può sopravvivere integralmente, sopravvivere frammentariamente (ma ancora leggibile almeno in parte), o essere completamente perduta. Possiamo considerare diversi endpoints: il momento della prima frammentazione, il momento della perdita completa del testo, il momento della rimozione dal contesto originario. La scelta dell'endpoint dipende dalla domanda di ricerca.

Un secondo problema metodologico riguarda la censura. In survival analysis, la censura si verifica quando non osserviamo l'evento di interesse per alcuni soggetti. Nel nostro caso, tutte le iscrizioni che ci sono pervenute sono "censurate a destra": non abbiamo osservato la loro distruzione perché sono ancora esistenti. Questo crea un bias potenziale: le iscrizioni sopravvissute possono non essere rappresentative di tutte le iscrizioni mai create. Tecniche standard di survival analysis come il metodo di Kaplan-Meier e i modelli di Cox proportional hazards sono specificamente progettati per gestire la censura.

Quali fattori influenzano la sopravvivenza delle iscrizioni? L'analisi empirica rivela diversi predittori significativi. Il materiale è cruciale: iscrizioni in marmo duro (come il marmo di Carrara o il granito egiziano) hanno tassi di sopravvivenza molto superiori a quelle in calcare tenero o arenaria. La dimensione conta: monumenti massicci hanno maggiore probabilità di sopravvivere di piccole stele. Il tipo di installazione è rilevante: iscrizioni murate in edifici o incorporate in fondazioni tendono a sopravvivere meglio di quelle free-standing esposte agli elementi.

Il contesto geografico e culturale influenza fortemente la sopravvivenza. Iscrizioni in regioni che subirono continuità di insediamento urbano (come Roma, dove l'interesse antiquario rinascimentale portò alla conservazione di molte iscrizioni) hanno tassi di sopravvivenza superiori a quelle in aree abbandonate o rurali. Paradossalmente, iscrizioni in regioni che furono completamente abbandonate dopo l'antichità (come alcune zone desertiche del Nordafrica o del Medio Oriente) talvolta mostrano eccellente conservazione proprio perché non furono esposte a secoli di riutilizzo edilizio.

Il contenuto dell'iscrizione può influenzare la sopravvivenza attraverso meccanismi culturali. Iscrizioni con contenuto cristiano avevano maggiore probabilità di essere preservate in contesti cristiani medievali, mentre iscrizioni pagane potevano essere distrutte per motivi religiosi. Iscrizioni che celebravano imperatori poi dannati dalla memoria (damnatio memoriae) furono talvolta deliberatamente cancellate o riutilizzate. Iscrizioni in materiali pregiati (come il porfido o il bronzo) erano particolarmente vulnerabili al riutilizzo.

Modelli di Cox proportional hazards permettono di quantificare simultaneamente l'effetto di multiple covariate sul rischio di distruzione. La hazard ratio associata a una covariata indica di quanto aumenta o diminuisce il rischio di distruzione quando quella covariata aumenta di una unità. Ad esempio, un'analisi potrebbe rivelare che essere realizzata in marmo piuttosto che in calcare riduce il rischio di distruzione di un fattore tre (hazard ratio = 0,33), mentre ogni metro cubo di volume aggiuntivo riduce il rischio del dieci per cento.

Un'applicazione interessante di questo approccio riguarda la stima del numero totale di iscrizioni originariamente create. Se possiamo modellare la probabilità di sopravvivenza in funzione di caratteristiche osservabili, possiamo in principio invertire il modello per stimare quante iscrizioni con certe caratteristiche furono originariamente create dato il numero di sopravvissute. Ad esempio, se stimiamo che iscrizioni funerarie in calcare del secondo secolo hanno una probabilità di sopravvivenza del due per cento, e ne osserviamo duemila, possiamo stimare che ne furono originariamente create circa centomila. Queste stime sono ovviamente molto incerte ma forniscono ordini di grandezza utili.

L'analisi della sopravvivenza differenziale ha implicazioni importanti per l'interpretazione dei dati epigrafici. Se certi tipi di iscrizioni sopravvivono sistematicamente meglio di altri, il record epigrafico che osserviamo sarà distorto. Ad esempio, se iscrizioni monumentali di élite in marmo sopravvivono molto meglio di semplici stele di gente comune in materiale deperibile, il nostro corpus epigrafico sovra-rappresenterà le élites. Correggere per questi bias di sopravvivenza è metodologicamente complesso ma concettualmente necessario per inferenze demografiche e sociali accurate.

Analisi linguistica: l'evoluzione del latino attraverso le "imperfezioni"

Le iscrizioni latine, specialmente quelle di carattere più privato e meno monumentale, preservano forme linguistiche che si discostano dal latino classico letterario e riflettono invece il latino realmente parlato nelle diverse regioni e periodi dell'impero. Questi "errori" ortografici, morfologici e sintattici dal punto di vista del latino normativo sono in realtà preziose testimonianze dell'evoluzione della lingua latina verso le lingue romanze, e la loro analisi quantitativa permette di tracciare processi di cambiamento linguistico con precisione cronologica e geografica impossibile da altre fonti.

Il latino epigrafico "substandard" presenta numerose categorie di deviazioni dalla norma classica. A livello fonologico, le confusioni grafiche rivelano cambiamenti nella pronuncia: la confusione tra E ed I in sillabe atone riflette il processo di indebolimento vocalico che porterà alla riduzione del sistema vocalico romanzo; la confusione tra O ed U riflette un processo parallelo nelle vocali posteriori; la perdita di H riflette la sua sparizione come fonema; la semplificazione di gruppi consonantici riflette tendenze fonotattiche evolutive. Ogni "errore" è in realtà una finestra sulla fonologia del latino parlato.

L'analisi quantitativa permette di tracciare quando e dove questi cambiamenti si verificano. Costruendo un database di forme "deviate" estratte dalle iscrizioni, georeferenziate e datate, è possibile mappare la diffusione geografica e temporale di ciascun fenomeno. Ad esempio, la confusione E/I appare sporadicamente già nel primo secolo, diventa progressivamente più frequente nel secondo e terzo secolo, ed è quasi universale nel quarto e quinto secolo. Ma la cronologia esatta varia geograficamente: alcune province mostrano il fenomeno prima di altre, suggerendo possibili centri di innovazione linguistica e percorsi di diffusione.

Le deviazioni morfologiche sono altrettanto interessanti. La confusione tra casi grammaticali riflette la progressiva erosione del sistema declinazionale latino che porterà alla sua completa scomparsa nelle lingue romanze. Le iscrizioni mostrano crescente confusione tra

accusativo e ablativo, tra genitivo e dativo, tra nominativo e accusativo. Quantificando la frequenza di queste confusioni nel tempo, è possibile datare con precisione il declino del sistema dei casi. Anche qui emergono variazioni geografiche: alcune regioni (come la Sardegna) sembrano conservare il sistema declinazionale più a lungo, mentre altre (come l'Africa) mostrano un collasso più precoce.

I fenomeni sintattici evolutivi sono più difficili da catturare nelle brevi iscrizioni, ma comunque presenti. L'emergere di costruzioni preposizionali in contesti dove il latino classico userebbe casi semplici (ad esempio "de + ablativo" al posto del genitivo) segnala la transizione verso i sistemi preposizionali delle lingue romanze. L'ordine delle parole mostra tendenze verso strutture SVO (soggetto-verbo-oggetto) più rigide rispetto alla relativa libertà del latino classico. L'uso crescente di participi passati in costruzioni perifrastiche anticipa i futuri perfetti composti romanzi.

L'analisi lessicale rivela l'emergere di vocabolario popolare e regionale. Termini assenti dalla letteratura classica ma che diventeranno centrali nelle lingue romanze appaiono nelle iscrizioni: "casa" al posto di "domus", "testa" al posto di "caput", forme verbali innovative. L'analisi quantitativa della frequenza lessicale nelle iscrizioni di diversi periodi e regioni permette di tracciare l'espansione di questi vocaboli innovativi.

Un aspetto metodologicamente importante riguarda la distinzione tra errori genuini che riflettono la lingua parlata ed errori puramente grafici dovuti all'incompetenza dell'incisore. Gli errori che riflettono cambiamenti fonologici reali dovrebbero mostrare patterns sistematici: se la confusione E/I riflette davvero una neutralizzazione fonologica, dovremmo osservarla consistentemente attraverso molte iscrizioni e molti contesti. Errori idiosincratici casuali, invece, suggeriscono incompetenza piuttosto che cambiamento linguistico. L'analisi statistica può distinguere patterns sistematici da noise casuale.

La stratificazione sociale influenza fortemente il tipo di latino utilizzato nelle iscrizioni. Iscrizioni di élite, redatte da scriptores professionali, tendono a conformarsi molto più strettamente alle norme classiche. Iscrizioni di gente comune, incise da lapicidi di scarsa formazione o dagli stessi committenti, mostrano molto più frequentemente forme popolari. Questa stratificazione sociale del latino epigrafico permette di studiare la sociolinguistica del mondo romano: quanto erano distanti la lingua dell'élite e quella popolare? Come variava questa distanza nel tempo e nello spazio?

Tecniche di clustering possono identificare automaticamente gruppi di iscrizioni linguisticamente affini, potenzialmente rivelando varietà regionali del latino. Un'iscrizione dalla Lusitania potrebbe mostrare features fonologiche che la accomunano ad altre iscrizioni lusitane ma la distinguono da iscrizioni galliche o africane, rivelando l'emergere di quella che diventerà la varietà iberica del latino da cui evolverà il portoghese. Queste analisi permettono di tracciare le radici della diversificazione romanza molto prima della frammentazione politica dell'impero.

Diffusione geografica: tracciare la propagazione culturale

Oltre all'analisi statica della distribuzione geografica, l'epigrafia quantitativa può tracciare processi dinamici di diffusione spaziale di pratiche culturali, culti religiosi, formule epigrafiche, e altri fenomeni culturali. Quando una nuova formula dedicatoria, un nuovo culto, un nuovo tipo di monumento funerario appare prima in una località e poi si diffonde in altre, questo processo lascia signature spaziali e temporali nel record epigrafico che possono essere analizzate quantitativamente.

I modelli di diffusione spaziale distinguono generalmente tra diffusione per contagio (dove un fenomeno si diffonde da luogo a luogo adiacente, come un'epidemia) e diffusione gerarchica (dove un fenomeno salta da un centro urbano importante a un altro, indipendentemente dalla distanza geografica). L'analisi epigrafica può rivelare quale tipo di diffusione caratterizza diversi fenomeni culturali romani.

Il culto di Giove Dolicheno, divinità siriana che conobbe grande diffusione nell'esercito romano del secondo e terzo secolo, mostra un pattern di diffusione chiaramente tracciabile attraverso le iscrizioni dedicatorie. Le prime attestazioni si concentrano nella Siria e nelle province orientali. Nel secondo secolo, il culto appare nelle guarnigioni militari della frontiera danubiana, poi in quelle renane, poi in Britannia. Mappando cronologicamente queste attestazioni, emerge un chiaro pattern di diffusione lungo le rotte di trasferimento delle unità militari. I soldati che avevano servito in Oriente portavano il culto con sé quando venivano trasferiti in altre province. Questo è un esempio di diffusione gerarchica mediata dalla rete militare piuttosto che da prossimità geografica.

Al contrario, certe formule epigrafiche mostrano diffusione per contagio. Una particolare formula commemorativa funeraria potrebbe apparire prima in una città e poi diffondersi gradualmente alle città vicine, creando un'onda di adozione che si propaga geograficamente. Modelli matematici di diffusione spaziale, come le equazioni di reazione-diffusione derivate dall'ecologia, possono essere adattati per modellare questi processi. Il parametro di diffusione D misura quanto rapidamente la pratica si diffonde nello spazio, mentre il parametro di crescita r misura quanto rapidamente viene adottata localmente.

L'analisi può anche identificare barriere alla diffusione. Confini provinciali, catene montuose, zone linguistiche diverse possono rallentare o bloccare la diffusione di pratiche culturali. Ad esempio, la diffusione di certe pratiche funerarie tipicamente occidentali si arresta nettamente al confine tra province di lingua latina e province di lingua greca, suggerendo che la barriera linguistica costituiva anche una barriera culturale significativa.

La velocità di diffusione può essere quantificata misurando il tempo necessario perché una pratica attraversi una certa distanza geografica. Se una formula appare a Roma nell'anno 50 d.C. e a Lione nell'anno 80 d.C., possiamo calcolare una velocità di diffusione di circa trenta anni per coprire la distanza Roma-Lione (circa 900 km), ovvero circa 30 km all'anno. Questa velocità può essere confrontata con velocità note di altri processi di diffusione culturale o epidemiologica per capire quanto rapidamente le idee e le pratiche circolavano nell'impero romano.

Un'applicazione particolarmente interessante riguarda la diffusione dei nomi personali imperiali attraverso la popolazione. Quando un nuovo imperatore saliva al trono e adottava certi cognomina, questi nomi tendevano a diffondersi rapidamente nella popolazione, sia per

emulazione sia perché nuovi cittadini che ricevevano la cittadinanza assumevano il nome gentilizio dell'imperatore regnante. Tracciando la frequenza di questi nomi nelle iscrizioni datate di diverse province, è possibile misurare la velocità con cui le mode onomastiche imperiali si diffondevano dalla capitale alle periferie.

I patterns di diffusione possono anche rivelare reti di contatto e influenza culturale non documentate da altre fonti. Se due città remote mostrano simultaneamente l'adozione della stessa pratica epigrafica innovativa, questo suggerisce l'esistenza di contatti diretti tra di loro, forse commerciali o attraverso migrazioni. L'analisi di reti di similarità culturale costruite sulla base di features epigrafiche condivise può quindi rivelare reti di interazione che complementano le reti commerciali e politiche documentate archeologicamente.

Modellazione bayesiana: stimare i dati mancanti

Una delle sfide fondamentali dell'epigrafia quantitativa è che il corpus di iscrizioni sopravvissute rappresenta solo una frazione delle iscrizioni originariamente create, e questa frazione è sistematicamente distorta da processi di conservazione differenziale. Le iscrizioni monumentali in marmo sopravvivono meglio delle semplici stele in calcare, le iscrizioni urbane sopravvivono meglio di quelle rurali, le iscrizioni di regioni con continuità insediativa sopravvivono meglio di quelle di aree abbandonate. Per fare inferenze sulla realtà storica passata piuttosto che sui processi di conservazione archeologica, è necessario correggere per questi bias e stimare le caratteristiche del corpus originale completo.

L'approccio bayesiano fornisce un framework formale per questo tipo di inferenza in presenza di dati incompleti e incerti. L'idea fondamentale è di specificare un modello probabilistico che descriva sia il processo generativo delle iscrizioni (quante e che tipo di iscrizioni venivano create in diversi periodi e luoghi) sia il processo di conservazione (quale frazione di iscrizioni di diversi tipi sopravvive fino al presente). Combinando questo modello con i dati osservati attraverso il teorema di Bayes, si ottiene una distribuzione posteriore sui parametri del modello che rappresenta formalmente la nostra incertezza su aspetti del passato non direttamente osservabili.

Un esempio concreto riguarda la stima della popolazione antica basata sulle iscrizioni funerarie. Supponiamo di voler stimare la popolazione di una città antica in un certo periodo basandoci sul numero di epitaffi datati a quel periodo. Il numero osservato di epitaffi N_{oss} dipende da tre fattori: la popolazione reale P , il tasso di commemorazione epigrafica c (quale frazione di morti veniva commemorata con un'iscrizione), e il tasso di sopravvivenza s (quale frazione di iscrizioni è sopravvissuta). Quindi $N_{oss} = P \times m \times c \times s$, dove m è il tasso di mortalità annuo.

Il problema è che non conosciamo né c né s con precisione. Tuttavia, possiamo specificare distribuzioni di probabilità a priori su questi parametri basate su conoscenze ausiliarie. Ad esempio, sappiamo da altre fonti che solo una frazione della popolazione (probabilmente tra il cinque e il venti per cento) poteva permettersi di commissionare epitaffi. Possiamo specificare una distribuzione di probabilità a priori su c che rifletta questa conoscenza. Similmente, basandoci su studi di conservazione archeologica, possiamo specificare una distribuzione a priori su s .

Utilizzando metodi di inferenza bayesiana come Markov Chain Monte Carlo (MCMC), possiamo campionare dalla distribuzione posteriore di P dato N_{oss} e ottenere non solo una stima puntuale della popolazione ma anche intervalli di credibilità che quantificano l'incertezza. Questo approccio è molto più onesto intellettualmente rispetto a stime puntuali che nascondono l'enorme incertezza inerente al problema.

Un'altra applicazione importante riguarda la ricostruzione di testi lacunosi. Molte iscrizioni ci sono pervenute frammentarie, con porzioni mancanti di testo. Gli epigrafisti propongono integrazioni basate su paralleli formulari e su vincoli spaziali (lunghezza della lacuna), ma queste integrazioni sono spesso multiple e incerte. L'approccio bayesiano permette di assegnare probabilità a diverse integrazioni possibili.

Si può costruire un modello linguistico statistico (ad esempio un modello n -gram o un modello neurale di linguaggio) addestrato sul corpus di iscrizioni complete. Questo modello assegna probabilità a sequenze di parole latine. Data un'iscrizione frammentaria, il modello può generare le sequenze più probabili che riempirebbero la lacuna, assegnando a ciascuna una probabilità. Invece di scegliere una singola integrazione "corretta", possiamo mantenere una distribuzione di probabilità sulle integrazioni possibili e propagare questa incertezza attraverso analisi successive.

Power-law distributions: ricchezza e disuguaglianza nell'impero

L'analisi della distribuzione statistica di variabili economiche estratte dalle iscrizioni può rivelare patterns di disuguaglianza economica e distribuzione della ricchezza nel mondo romano. Molte iscrizioni forniscono informazioni quantitative sui costi dei monumenti, sulle donazioni evergete, sui patrimoni dichiarati, sui salari e prezzi. Quando aggregiamo queste informazioni da migliaia di iscrizioni, emergono distribuzioni statistiche che possono essere caratterizzate formalmente e che rivelano la struttura economica della società romana.

Una delle osservazioni più notevoli riguarda la distribuzione dei costi dei monumenti funerari. Alcune iscrizioni, specialmente quelle commissionate da collegia funerari o da associazioni professionali, riportano esplicitamente il costo del monumento in sesterzi. Raccogliendo queste informazioni da centinaia di iscrizioni, si può costruire una distribuzione empirica dei costi. Questa distribuzione risulta essere estremamente skewed (asimmetrica): la maggior parte dei monumenti costa poche migliaia di sesterzi, ma esiste una lunga coda di monumenti progressivamente più costosi, fino a monumenti di élite che costano centinaia di migliaia o addirittura milioni di sesterzi.

Questa distribuzione approssima una power-law (legge di potenza), dove la probabilità $P(x)$ che un monumento costi x è proporzionale a x elevato a un esponente negativo α : $P(x) \propto x^{(-\alpha)}$. Le distribuzioni power-law sono caratterizzate dalla proprietà di scale invariance: non esiste una scala tipica, e la distribuzione ha la stessa forma quando osservata a diverse scale. Sono anche caratterizzate da "code pesanti": eventi estremi (monumenti estremamente costosi) sono molto più probabili di quanto sarebbero in una distribuzione normale.

L'esponente α della power-law fornisce una misura quantitativa della disuguaglianza economica. Valori più bassi di α indicano maggiore disuguaglianza (la coda pesante è ancora più pesante), mentre valori più alti indicano distribuzione più equa. Stimare α dai dati richiede tecniche statistiche appropriate, perché metodi standard come la regressione lineare sul log-log plot producono stime distorte. Il metodo della maximum likelihood è preferibile, possibilmente complementato da test di goodness-of-fit per verificare che la power-law sia effettivamente un buon modello per i dati.

L'analisi delle distribuzioni di ricchezza può essere condotta geograficamente e temporalmente. Confrontando l'esponente α in diverse province, si può quantificare se alcune regioni erano più o meno inegualitarie. L'Africa proconsolare, ad esempio, potrebbe mostrare maggiore disuguaglianza dell'Italia, riflettendo la concentrazione di enormi latifondi in mani di pochi proprietari terrieri. Tracciando l'evoluzione di α nel tempo, si può vedere se la disuguaglianza aumentava o diminuiva attraverso i secoli dell'impero.

Le donazioni evergete riportate nelle iscrizioni onorarie forniscono un altro dataset per analisi di power-law. Evergeti facoltosi donavano somme alle loro città per costruire edifici pubblici, finanziare giochi, distribuire grano, costruire acquedotti. Le iscrizioni commemorative spesso riportano le somme donate. Anche queste donazioni seguono distribuzioni power-law: molti evergeti donavano somme moderate (decine o centinaia di migliaia di sesterzi), pochi donavano somme enormi (milioni di sesterzi). La forma di questa distribuzione riflette la struttura della ricchezza dell'élite municipale.

Un aspetto interessante riguarda la comparazione tra diverse società antiche. Possiamo costruire distribuzioni di ricchezza da iscrizioni greche, romane, ed egiziane e confrontare i loro esponenti power-law. Questo fornisce una misura quantitativa comparativa di disuguaglianza economica cross-culturale. È possibile, ad esempio, che la società romana fosse quantificabilmente più diseguale della società greca classica, ma meno diseguale di alcune monarchie ellenistiche.

Le distribuzioni power-law appaiono anche in altri contesti epigrafici. La distribuzione della "produttività epigrafica" degli individui (quante volte un individuo appare menzionato in iscrizioni) segue una power-law: pochi individui (tipicamente imperatori e membri dell'élite senatoria) appaiono in centinaia o migliaia di iscrizioni, mentre la grande maggioranza della popolazione appare in una o poche iscrizioni. Questo riflette la distribuzione power-law della rilevanza sociale e del potere.

Anche il numero di iscrizioni per località segue approssimativamente una power-law: poche grandi città (Roma, Ostia, Pozzuoli) hanno decine di migliaia di iscrizioni, molte cittadine medie hanno centinaia o migliaia di iscrizioni, moltissimi piccoli insediamenti hanno decine o poche iscrizioni. Questa distribuzione riflette la gerarchia urbana dell'impero e segue approssimativamente la relazione rank-size tipica dei sistemi urbani.

L'interpretazione teorica delle power-laws osservate richiede attenzione. Le distribuzioni power-law possono emergere da molti processi generativi diversi, quindi osservare una power-law non identifica automaticamente il meccanismo sottostante. Tuttavia, in contesti economici, le power-laws sono spesso associate a processi di crescita moltiplicativa con preferential attachment: i ricchi diventano più ricchi proporzionalmente a quanto sono già

ricchi. Questo meccanismo, noto come effetto Matteo ("a chi ha sarà dato"), può generare e mantenere disuguaglianze estreme.

Epigrafia comparativa: analisi cross-culturale quantitativa

L'epigrafia quantitativa permette confronti sistematici tra diverse culture epigrafiche antiche: greca, romana, egiziana geroglifica e demotica, fenicia, aramaica, nabatea, palmirena. Questi confronti possono essere formulati in termini quantitativi piuttosto che puramente qualitativi, permettendo di testare ipotesi sulle differenze culturali in modo statisticamente rigoroso.

Un primo livello di comparazione riguarda semplicemente la densità epigrafica: quante iscrizioni per capita o per unità di area producevano diverse culture? Questa è una misura della "epigrafia-intensità" di una cultura. La cultura romana appare straordinariamente epigrafia-intensiva: le province romane producevano iscrizioni a un tasso molto superiore rispetto a culture contemporanee o precedenti. La Grecia classica, pur avendo una ricca tradizione epigrafica, produceva meno iscrizioni pro-capite di quanto facesse una provincia romana media. Questa differenza quantitativa riflette probabilmente differenze nella struttura sociale: la società romana era più municipale, più competitiva status-wise, più orientata alla commemorazione pubblica.

Un secondo livello di comparazione riguarda i contenuti tipologici. Possiamo quantificare quale percentuale delle iscrizioni in diverse culture appartiene a diverse categorie: funerarie, dedicatorie, onorarie, editti, leggi, trattati, ecc. La cultura romana mostra una preponderanza di iscrizioni funerarie (forse settanta percento del totale), riflettendo una forte enfasi sulla commemorazione individuale e familiare. La cultura greca classica mostra una proporzione maggiore di iscrizioni pubbliche e decreti, riflettendo forse un'enfasi maggiore sulla polis rispetto alla famiglia. L'epigrafia egiziana templare è dominata da testi religiosi e commemorazioni regali.

Le distribuzioni onomastiche differiscono drammaticamente tra culture e possono essere confrontate quantitativamente. Il sistema onomastico romano con i suoi tria nomina è estremamente distintivo e può essere quantificato in termini di entropia informativa: quanta informazione sociale codifica un nome? Un nome romano completo con praenomen, nomen, cognomen, filiazione, e tribù codifica informazioni sulla famiglia paterna, sulla tribù di registrazione (e quindi approssimativamente sull'origine geografica), e attraverso il cognomen potenzialmente sull'origine etnica o su caratteristiche familiari. Questo può essere confrontato quantitativamente con sistemi onomastici più semplici.

Le formule standardizzate differiscono tra culture. L'epigrafia romana fa uso massiccio di abbreviazioni standardizzate (DM per Dis Manibus, HSE per Hic Situs Est, ecc.), creando un alto grado di compressione testuale. Possiamo quantificare il "tasso di abbreviazione" come la frazione di parole abbreviate rispetto al totale, e confrontarlo tra culture. L'epigrafia greca usa meno abbreviazioni dell'epigrafia latina. L'epigrafia geroglifica egiziana, pur usando un sistema di scrittura radicalmente diverso, mostra anch'essa alto grado di formulizzazione e può essere analizzata in termini di ripetizione di sequenze standard.

L'analisi stilometrica comparativa può quantificare differenze nello stile testuale tra culture. Misure come la complessità sintattica, la lunghezza media delle frasi, la ricchezza lessicale (rapporto type/token), la distribuzione delle classi grammaticali possono essere calcolate per corpora epigrafici di diverse culture e confrontate. L'epigrafia latina tende ad essere sintatticamente più compressa e ellittica dell'epigrafia greca, riflettendo forse differenze tra le lingue ma anche differenze nelle convenzioni stilistiche epigrafiche.

I patterns cronologici della produzione epigrafica differiscono significativamente. Mentre l'epigrafia romana mostra il caratteristico picco antonino del secondo secolo seguito da declino nel terzo e quarto secolo, l'epigrafia greca mostra un pattern diverso, con produzione sostenuta attraverso il periodo imperiale. L'epigrafia geroglifica egiziana mostra una lunghissima tradizione che attraversa millenni con variazioni di intensità legate alle dinastie e ai periodi di centralizzazione politica. Confrontare quantitativamente questi patterns temporali illumina le diverse traiettorie culturali.

I patterns spaziali differiscono tra culture in modi quantificabili. L'epigrafia romana mostra forte concentrazione urbana, con la grande maggioranza delle iscrizioni provenienti da contesti urbani o peri-urbani. L'epigrafia greca mostra pattern simili ma con maggiore presenza di iscrizioni nei santuari extra-urbani. L'epigrafia nabatea e palmirena mostra concentrazione in certi centri culturali e commerciali del deserto. Misure di clustering spaziale possono quantificare queste differenze.

Un aspetto particolarmente interessante riguarda la distribuzione dei topics testuali. Applicando topic modeling a corpora epigrafici di diverse culture (ovviamente dopo aver tradotto o trascritto i testi in forme comparabili), possiamo identificare topics caratteristici di ciascuna cultura e topics condivisi. Questo può rivelare continuità e discontinuità culturali. Ad esempio, il topic della famiglia nucleare è prominente nell'epigrafia romana ma meno in quella greca, dove prevalgono riferimenti alla polis e alle sue istituzioni.

Machine learning per la ricostruzione: predire testo lacunoso

Una delle applicazioni più ambiziose del machine learning all'epigrafia riguarda la ricostruzione automatica di porzioni mancanti di testo in iscrizioni frammentarie. Questo è un problema di sequence-to-sequence prediction: data una sequenza di parole con lacune, predire quale sequenza di parole riempie le lacune. Recenti progressi nel deep learning, particolarmente con modelli transformer come BERT e GPT, hanno aperto nuove possibilità per questo tipo di task.

Il problema può essere formulato come masked language modeling: si addestra un modello neurale su un grande corpus di iscrizioni complete, dove casualmente si mascherano alcune parole e si chiede al modello di predirle basandosi sul contesto circostante. Una volta addestrato, il modello può essere applicato a lacune reali in iscrizioni frammentarie. Il modello impara implicitamente le regolarità formulari, sintattiche e lessicali del latino epigrafico e le utilizza per fare predizioni plausibili.

I modelli transformer sono particolarmente adatti per questo task perché possono catturare dipendenze a lungo raggio nel testo. In un'iscrizione funeraria, la presenza di certe formule iniziali può vincolare fortemente quali formule appariranno verso la fine, anche se sono separate da molte parole. Il meccanismo di attention dei transformer permette al modello di "guardare" simultaneamente a tutto il contesto disponibile quando fa una predizione.

Un vantaggio significativo dell'approccio di machine learning rispetto alle tradizionali integrazioni filologiche è che il modello può fornire non solo la singola integrazione più probabile, ma una distribuzione di probabilità su multiple integrazioni possibili, ordinata per plausibilità. Questo permette all'epigrafo di considerare alternative e di valutare l'incertezza della ricostruzione. Se il modello assegna probabilità quasi uguale a due diverse integrazioni, questo segnala genuina ambiguità che non può essere risolta dal contesto testuale.

La qualità delle ricostruzioni dipende criticamente dalla quantità e qualità dei dati di training. Per il latino epigrafico, abbiamo il vantaggio di centinaia di migliaia di iscrizioni complete che possono essere utilizzate per training. Tuttavia, il latino epigrafico mostra significativa variazione nel tempo, nello spazio, e tra tipologie, quindi idealmente il modello dovrebbe essere addestrato su iscrizioni comparabili all'iscrizione frammentaria che si vuole ricostruire. Approcci di transfer learning permettono di addestrare prima un modello generale su tutto il corpus e poi fine-tunarlo su sottoinsiemi specifici.

Un problema metodologico importante riguarda la valutazione. Come possiamo misurare l'accuratezza delle ricostruzioni automatiche? Un approccio consiste nel creare artificialmente iscrizioni frammentarie a partire da iscrizioni complete nel dataset di test: si rimuovono deliberatamente porzioni di testo, si chiede al modello di ricostruirle, e si confronta la ricostruzione con il testo originale noto. Metriche come l'accuratezza a livello di parola, il BLEU score (usato in machine translation), o misure di similarità semantica possono quantificare la performance.

Le valutazioni empiriche mostrano che i modelli transformer moderni possono ricostruire correttamente circa il cinquanta-settanta per cento delle parole mancanti in lacune brevi (una-tre parole) in iscrizioni standardizzate e formulari, ma la performance degrada significativamente per lacune più lunghe o per iscrizioni meno formulari. Tuttavia, anche quando la predizione esatta non è corretta, le predizioni del modello sono spesso semanticamente appropriate e formulaicamente plausibili, suggerendo che il modello ha effettivamente appreso le convenzioni del genere epigrafico.

Un'estensione interessante riguarda l'integrazione di informazioni non testuali. Molte lacune in iscrizioni hanno lunghezza fisica nota (si può misurare lo spazio disponibile sulla pietra), e questa informazione vincola il numero di lettere che la lacuna può contenere. I modelli possono essere modificati per incorporare questo vincolo, escludendo ricostruzioni che sarebbero troppo lunghe o troppo corte per lo spazio disponibile. Similmente, informazioni paleografiche (tracce di lettere parzialmente conservate) possono essere incorporate come soft constraints sulle predizioni.

L'approccio di machine learning alla ricostruzione epigrafica non mira a sostituire l'expertise filologica umana, ma piuttosto a assisterla. Il modello può rapidamente generare integrazioni candidate plausibili che l'epigrafo può poi valutare criticamente. In casi ambigui, il modello

può suggerire alternative che l'epigrafo potrebbe non aver considerato. In grandi progetti di edizione digitale, il modello può fornire integrazioni preliminari per migliaia di frammenti, accelerando significativamente il lavoro editoriale.

Scansione 3D e database mesh: la materialità digitale

L'epigrafia quantitativa non si limita all'analisi del testo, ma sempre più incorpora la documentazione e l'analisi della materialità delle iscrizioni attraverso tecnologie di acquisizione 3D. La fotogrammetria, la scansione laser, e la structured light scanning permettono di creare modelli tridimensionali digitali accurati di monumenti epigrafici, preservando informazioni sulla forma, le dimensioni, lo stato di conservazione, e le caratteristiche paleografiche che sarebbero impossibili da catturare attraverso fotografie bidimensionali o disegni.

La fotogrammetria structure-from-motion (SfM) è diventata particolarmente popolare per la documentazione epigrafica perché richiede solo una fotocamera standard e software di processing. Il metodo funziona acquisendo molte fotografie dello stesso oggetto da angolazioni diverse, poi utilizzando algoritmi di computer vision per identificare features corrispondenti tra le immagini e triangolare la posizione 3D di ogni punto sulla superficie dell'oggetto. Il risultato è una nuvola di punti tridimensionale e una mesh poligonale texture-mapped che rappresenta fedelmente la geometria e l'aspetto dell'iscrizione.

I modelli 3D di iscrizioni offrono numerosi vantaggi per la ricerca. Permettono misurazioni accurate di dimensioni, angoli, profondità di incisione, che sarebbero difficili o impossibili da ottenere sul manufatto fisico, specialmente se questo è montato in alto o in posizione inaccessibile. Permettono visualizzazioni sotto illuminazione virtuale da qualsiasi angolo, spesso rivelando dettagli di lettere erose che sono invisibili a occhio nudo sotto illuminazione normale. Permettono la creazione di rappresentazioni derivate come disegni a filo, mappe di profondità, o visualizzazioni con illuminazione raking virtuale.

Una tecnica particolarmente utile è il Reflectance Transformation Imaging (RTI) o Polynomial Texture Mapping (PTM), che cattura l'oggetto sotto illuminazione da molte direzioni diverse e crea una rappresentazione matematica di come ogni pixel riflette la luce in funzione della direzione di illuminazione. Questo permette al ricercatore di "ri-illuminare" virtualmente l'iscrizione da qualsiasi direzione desiderata, evidenziando dettagli superficiali minimali che rivelano lettere altrimenti invisibili. L'RTI si è rivelato particolarmente efficace per leggere iscrizioni erose, graffiti leggerissimi, o palinsesti dove un testo è stato parzialmente cancellato e sovrascritto.

La costruzione di database di modelli 3D di iscrizioni permette nuove forme di analisi quantitativa. È possibile estrarre automaticamente misure morfometriche da migliaia di iscrizioni: altezza e larghezza delle lettere, spacing tra lettere e tra righe, profondità di incisione, angolo di incisione. Queste misure possono poi essere analizzate statisticamente per identificare patterns temporali, geografici, o associati a specifiche officine lapidarie. Ad esempio, la profondità media di incisione decresce nel tempo, riflettendo forse cambiamenti nelle tecniche o negli strumenti.

L'analisi paleografica computazionale può essere significativamente migliorata utilizzando dati 3D piuttosto che immagini 2D. La profondità e la sezione del solco delle lettere contengono informazioni importanti sullo stile e sulla tecnica che non sono catturate da fotografie. Features 3D come la curvatura della superficie, la direzione dei gradienti di profondità, possono essere estratte e utilizzate come input per algoritmi di classificazione paleografica.

I modelli 3D permettono anche ricostruzioni virtuali di monumenti frammentari. Se due o più frammenti di una iscrizione sono conservati separatamente, i loro modelli 3D possono essere virtualmente riassemblati provando diverse configurazioni fino a trovare quella che minimizza i gap e massimizza la continuità delle features (linee di frattura, continuità del testo). Questo processo di virtual anastilosis è molto più efficiente e meno invasivo rispetto a tentativi fisici di riassembraggio.

La disseminazione di modelli 3D attraverso database online permette accesso democratizzato a iscrizioni che fisicamente sono conservate in musei remoti o in situ in località difficilmente accessibili. Progetti come Europeana, Arachne, o database specializzati come il Heidelberg Epigraphic Database 3D stanno creando repository di modelli 3D di iscrizioni che possono essere visualizzati, misurati, e analizzati da chiunque abbia una connessione internet. Questo trasforma radicalmente le possibilità di ricerca collaborativa e comparativa.

Un aspetto emergente riguarda l'uso di realtà virtuale e aumentata per l'esperienza delle iscrizioni. I modelli 3D possono essere visualizzati in ambienti VR immersivi, permettendo al ricercatore di "stare di fronte" all'iscrizione a dimensione reale e esaminarla da ogni angolo. La realtà aumentata permette di sovrapporre ricostruzioni virtuali di iscrizioni frammentarie o integrazioni di testo mancante direttamente sulla vista del frammento fisico, facilitando comprensione e interpretazione.

Network of stone: reti commerciali del marmo

Un'applicazione affascinante dell'epigrafia quantitativa combinata con l'archeometria riguarda la ricostruzione delle reti commerciali attraverso l'analisi della provenienza geologica dei materiali lapidei utilizzati per le iscrizioni. Il marmo e altri materiali litici pregiati utilizzati per iscrizioni monumentali venivano estratti da cave specifiche e poi trasportati, spesso su grandi distanze, ai luoghi di utilizzo. Identificando la provenienza geologica di un'iscrizione attraverso analisi petrografica o geochemica, e combinando questa informazione con la localizzazione del ritrovamento, è possibile tracciare flussi commerciali che collegavano cave, centri di lavorazione, e destinazioni finali.

L'identificazione della provenienza geologica si basa su caratteristiche petrografiche e geochemiche distintive di diversi marmi. Il marmo di Carrara (Luna) ha una grana caratteristica e composizione isotopica distintiva; il marmo pentelico da Atene contiene tipicamente mica moscovita che lo distingue da altri marmi bianchi; il marmo proconnesio dalla Propontide ha venature caratteristiche; i porfidi egiziani hanno composizioni mineralogiche uniche. Tecniche analitiche come la petrografia in sezione sottile, la spettrometria di massa per isotopi stabili (carbonio e ossigeno), la fluorescenza a raggi X per elementi in traccia, permettono di discriminare tra provenienze con alto grado di confidenza.

Quando migliaia di iscrizioni vengono analizzate per la loro provenienza geologica, emerge una complessa rete commerciale che può essere visualizzata e analizzata formalmente come un grafo. I nodi di questo grafo rappresentano località: le cave di estrazione da un lato, e le città o siti di ritrovamento dall'altro. Gli archi rappresentano flussi di materiale: se a Londinium (Londra romana) sono state trovate venti iscrizioni in marmo di Carrara, esiste un arco di peso venti dalla cava di Carrara a Londinium.

L'analisi di questa rete rivela patterns notevoli. Roma naturalmente funziona come hub centrale, importando materiali da tutte le principali cave dell'impero: marmo di Carrara dall'Italia settentrionale, marmo greco dalle Cicladi e dall'Attica, marmo dell'Asia Minore, porfido e granito dall'Egitto. Ma emerge anche che molte province importavano direttamente dalle cave senza passare necessariamente per Roma. Ad esempio, la Britannia importava marmo proconnesio direttamente via mare dal Mar di Marmara, bypassando l'Italia. La Spagna importava abbondantemente marmi nordafricani, riflettendo stretti legami commerciali transmediterranei.

Le misure di network centrality identificano i nodi più importanti. Roma ovviamente domina come principale destinazione. Ma emergono anche hub secondari: città portuali come Ostia, Aquileia, o Narbona che fungevano da centri di redistribuzione, ricevendo materiali grezzi o semi-lavorati e redistribuendoli a destinazioni interne. Le cave maggiori (Carrara, cave proconnesie, cave penteliche) emergono come sources centrali con collegamenti a molte destinazioni.

La struttura temporale della rete è altrettanto interessante. Tracciando l'evoluzione della rete attraverso i secoli, si osservano significativi cambiamenti. Nel primo secolo a.C. e primo secolo d.C., il commercio dei marmi è dominato da provenienze italiane e greche. Nel secondo secolo, si osserva un'espansione drammatica con l'integrazione di cave dall'Asia Minore, dall'Egitto, e dal Nordafrica. Questo riflette sia la crescente integrazione economica dell'impero sia probabilmente l'apertura di nuove rotte commerciali e lo sfruttamento di nuove cave. Nel terzo secolo, si osserva una certa contrazione con maggiore prevalenza di materiali locali, riflettendo forse difficoltà economiche o la frammentazione delle reti commerciali a lunga distanza.

La geografia della rete non è isotropica ma segue chiaramente le rotte marittime. Il trasporto marittimo era enormemente più efficiente ed economico del trasporto terrestre: trasportare un blocco di marmo di una tonnellata per mille chilometri via mare costava una frazione di quanto costasse trasportarlo per cento chilometri via terra. Di conseguenza, la rete dei marmi è essenzialmente una rete marittima, con le città costiere che fungono da nodi principali e le città interne che ricevono materiali principalmente dai porti vicini. Questa struttura è quantificabile: la probabilità che una città riceva marmi da una cava distante decresce con la distanza via mare ma è quasi indipendente dalla distanza terrestre dal porto più vicino.

L'analisi della rete permette anche di identificare anomalie interessanti che possono riflettere eventi storici o circostanze speciali. Ad esempio, un improvviso afflusso di porfido egiziano in una provincia periferica in un periodo specifico potrebbe riflettere un progetto edilizio imperiale o la presenza temporanea di un funzionario imperiale di alto rango. Cluster di

iscrizioni in un materiale raro tutti datati allo stesso periodo in una località specifica potrebbero indicare l'arrivo di un singolo grande carico o di un artigiano specializzato.

La network analysis può anche testare ipotesi sulla organizzazione del commercio dei marmi. Era un commercio diretto tra cave e consumatori finali, o passava attraverso intermediari e mercanti specializzati? L'esistenza di hub di redistribuzione suggerisce un sistema di intermediazione. La presenza di "isole" nella rete (sottoreti densamente connesse internamente ma poco connesse al resto) potrebbe riflettere circuiti commerciali regionali relativamente indipendenti.

Un aspetto metodologicamente sofisticato riguarda la gestione dell'incertezza nell'attribuzione della provenienza. Non tutte le attribuzioni petrochimiche sono ugualmente certe: alcuni marmi sono diagnostici e possono essere attribuiti con alta confidenza a una cava specifica, altri sono più ambigui e potrebbero provenire da diverse cave con geologie simili. Approcci di network inference bayesiani permettono di incorporare questa incertezza: invece di disegnare un singolo arco deterministico, si assegna una distribuzione di probabilità sulle possibili provenienze, e questa incertezza viene propagata attraverso l'analisi della rete.

La combinazione di dati epigrafici (datazione, tipologia, contesto sociale delle iscrizioni) e dati archeometrici (provenienza del materiale) permette analisi particolarmente ricche. Si può chiedere: cambiava la provenienza dei materiali in funzione del tipo di iscrizione? Le iscrizioni imperiali utilizzavano marmi diversi dalle iscrizioni private? Le iscrizioni funerarie di élite utilizzavano materiali più prestigiosi? Le risposte a queste domande illuminano sia l'organizzazione del commercio sia le gerarchie di prestigio sociale.

Integrazione di metodi: verso una epigrafia computazionale integrata

La frontiera attuale dell'epigrafia quantitativa non consiste tanto nello sviluppo di singole tecniche isolate, quanto piuttosto nella loro integrazione in workflow analitici complessi che combinano multiple metodologie per affrontare domande storiche sofisticate. Un esempio paradigmatico potrebbe essere uno studio sulla mobilità sociale nell'impero romano che combini analisi onomastica quantitativa, network analysis prosopografica, analisi spaziale della distribuzione delle iscrizioni, text mining per estrarre informazioni sulle carriere, e modellazione statistica per correggere bias di conservazione.

Il workflow potrebbe iniziare con l'estrazione automatica via NER di tutti i nomi personali dalle iscrizioni, insieme alle loro caratteristiche (filiazione, tribù, cariche, origine dichiarata). Questi dati alimenterebbero un'analisi onomastica che classifica gli individui secondo criteri sociali: cittadini romani di vecchia data versus nuovi cittadini (basandosi sui patterns onomastici), ingenui versus liberti (identificabili da marcatori onomastici specifici), élites versus non-élites (distinguibili dalle cariche ricoperte o dall'assenza di cariche).

Parallelamente, si costruirebbe una rete prosopografica codificando relazioni familiari, di patronato, e professionali estratte dalle stesse iscrizioni. Algoritmi di community detection identificherebbero clusters sociali, e misure di centralità identificherebbero individui

particolarmente influenti. L'analisi della struttura della rete potrebbe rivelare se la società romana era segmentata in gruppi sociali relativamente chiusi o se esistevano significative possibilità di mobilità tra strati sociali.

L'analisi spaziale mapperebbe geograficamente le carriere individuali tracciabili attraverso le iscrizioni: individui che iniziarono la carriera in una provincia e la conclusero in un'altra, soldati che migrarono dalla provincia di arruolamento attraverso varie guarnigioni, mercanti i cui affari li portarono attraverso l'impero. Pattern di mobilità geografica potrebbero essere correlati con mobilità sociale: gli individui più mobili geograficamente erano anche quelli che sperimentavano maggiore mobilità sociale?

Topic modeling applicato ai testi epigrafici potrebbe identificare discorsi di auto-presentazione caratteristici di diversi gruppi sociali. Come si presentavano epigraficamente i liberti arricchiti rispetto agli aristocratici di vecchia data? Quali virtù enfatizzavano? Quali aspetti delle loro carriere mettevano in evidenza? Le differenze quantificate nei profili testuali rivelerebbero diverse strategie di costruzione dell'identità sociale.

Modelli statistici multivariati integrerebbero tutte queste informazioni per testare ipotesi sulla mobilità sociale. Ad esempio, si potrebbe modellare la probabilità di raggiungere cariche di alto rango in funzione di: origine geografica, status iniziale (ingenuo/liberto), ricchezza (inferita dal costo dei monumenti), connessioni sociali (posizione nella rete prosopografica), periodo temporale, provincia di attività. I coefficienti di questo modello quantificherebbero l'importanza relativa di diversi fattori nel determinare il successo sociale.

Un aspetto cruciale di questo approccio integrato è la gestione sistematica dell'incertezza. Ogni componente dell'analisi introduce incertezza: l'estrazione automatica di nomi ha un certo tasso di errore, le datazioni sono spesso imprecise, le attribuzioni prosopografiche possono essere ambigue, la disambiguazione di individui con nomi identici è probabilistica. Un framework bayesiano permette di propagare formalmente questa incertezza attraverso tutte le fasi dell'analisi, producendo alla fine non stime puntuali ma distribuzioni di probabilità sui parametri di interesse.

La visualizzazione integrata dei risultati è essenziale per comunicare analisi così complesse. Dashboard interattive potrebbero permettere di esplorare simultaneamente la rete prosopografica, la distribuzione geografica delle iscrizioni, l'evoluzione temporale dei patterns, e le caratteristiche testuali, con tutte le visualizzazioni collegate in modo che selezionare elementi in una influenzi le altre. Tecniche di dimensionality reduction come t-SNE o UMAP potrebbero proiettare individui o iscrizioni in spazi bidimensionali dove la vicinanza rappresenta similarità multidimensionale, permettendo di identificare visivamente clusters e outliers.

Sfide metodologiche e direzioni future

Nonostante i progressi impressionanti, l'epigrafia quantitativa affronta ancora numerose sfide metodologiche significative. Una sfida fondamentale riguarda la qualità e la completezza dei dati. Anche i migliori database epigrafici contengono errori di trascrizione, metadati incompleti, datazioni incerte. La digitalizzazione non è uniforme: alcune regioni e periodi

sono molto meglio rappresentati di altri. Molte iscrizioni sono note solo da trascrizioni antiche e non più verificabili. La lacunosità e la frammentarietà sono pervasive.

Queste limitazioni dei dati pongono problemi per l'analisi quantitativa. Molte tecniche statistiche assumono dati completi o missing at random, ma i dati epigrafici sono missing not at random: le iscrizioni mancanti non sono un campione casuale di quelle originariamente create, ma sono sistematicamente distorte da processi di conservazione. Sviluppare metodi statistici robusti che possano produrre inferenze valide nonostante questi problemi di dati è una sfida metodologica centrale.

Una seconda sfida riguarda la validazione dei risultati. In molte analisi storiche quantitative, non esiste un ground truth rispetto al quale validare le conclusioni. Se un modello di machine learning predice che una certa iscrizione frammentaria conteneva una certa parola, come possiamo sapere se la predizione è corretta quando il testo originale è perduto per sempre? Tecniche di cross-validation e altre forme di validazione interna aiutano, ma non sostituiscono completamente la validazione esterna.

Una terza sfida riguarda l'interpretabilità dei modelli complessi. I modelli di deep learning possono raggiungere performance predittive impressionanti ma sono essenzialmente black boxes: è difficile capire perché fanno certe predizioni. Per la ricerca storica, comprendere il perché è spesso più importante della predizione stessa. C'è quindi tensione tra potenza predittiva e interpretabilità. Sviluppare metodi di explainable AI per applicazioni epigrafiche è un'area di ricerca attiva.

Una quarta sfida riguarda l'integrazione dell'expertise dominio-specifica con i metodi computazionali. L'epigrafia tradizionale ha accumulato secoli di conoscenze raffinate su abbreviazioni, formule, contesti archeologici, tradizioni regionali. Molto di questa conoscenza esiste nella forma di expertise tacita degli studiosi esperti piuttosto che in forma codificata ed esplicita. Tradurre questa expertise in forme utilizzabili computazionalmente (ad esempio, come constraints in modelli statistici o come features in algoritmi di machine learning) è difficile ma essenziale.

Guardando al futuro, diverse direzioni di sviluppo appaiono particolarmente promettenti. L'integrazione sempre più stretta con altre fonti di evidenza (archeologica, papirologia, numismatica) attraverso linked data e knowledge graphs potrebbe permettere analisi storiche veramente multimodali. L'applicazione di tecniche sempre più sofisticate di causal inference potrebbe permettere di andare oltre correlazioni per identificare relazioni causali tra fenomeni. Lo sviluppo di digital twins di siti antichi che integrano evidenza epigrafica, archeologica, e testuale in ricostruzioni virtuali immersive potrebbe trasformare la nostra capacità di comprendere i contesti originali delle iscrizioni.

L'intelligenza artificiale generativa (modelli come GPT o simili addestrati specificamente su testi epigrafici) potrebbe eventualmente assistere nella redazione di edizioni critiche, nella proposta di integrazioni, nell'identificazione di paralleli testuali. Sistemi di raccomandazione potrebbero suggerire agli studiosi iscrizioni rilevanti per le loro ricerche che altrimenti potrebbero non conoscere, scoprendo connessioni non ovvie attraverso l'enorme corpus digitalizzato.

La democratizzazione degli strumenti di analisi quantitativa attraverso interfacce user-friendly e piattaforme web-based potrebbe rendere queste tecniche accessibili non solo a specialisti computazionali ma alla più ampia comunità di epigrafisti e storici. Citizen science projects potrebbero coinvolgere il pubblico generale nella trascrizione, annotazione, e analisi di iscrizioni, moltiplicando esponenzialmente le risorse umane disponibili.

Considerazioni epistemologiche: numeri e interpretazione storica

L'adozione di metodi quantitativi in epigrafia solleva questioni epistemologiche profonde sul rapporto tra quantificazione e interpretazione storica. I numeri possono sembrare fornire una patina di oggettività e precisione, ma l'interpretazione storica rimane fondamentalmente un'attività interpretativa che richiede giudizio, contesto, e sensibilità alle complessità umane che non sempre si lasciano catturare da algoritmi.

Una prima questione riguarda il rapporto tra pattern quantitativi e significato storico. L'analisi quantitativa può rivelare che una certa formula epigrafica divenne più frequente nel secondo secolo, ma questa osservazione quantitativa non spiega automaticamente perché questo avvenne. La spiegazione richiede contestualizzazione storica: cambiamenti nelle pratiche sociali, nelle norme culturali, negli equilibri politici. I numeri possono suggerire domande, identificare patterns che richiedono spiegazione, ma non forniscono le spiegazioni stesse.

Una seconda questione riguarda il rischio di reificazione: trattare costrutti quantitativi (come un indice di centralità in una rete sociale, o un cluster identificato da un algoritmo) come se fossero entità storiche reali piuttosto che artefatti analitici. Le comunità identificate da algoritmi di community detection in reti prosopografiche potrebbero non corrispondere a gruppi sociali che gli attori storici riconoscevano come tali. I topics identificati da LDA sono costrutti statistici che massimizzano certe funzioni obiettivo, non necessariamente categorie semantiche naturali.

Una terza questione riguarda la relazione tra micro e macro, tra l'individuale e l'aggregato. L'epigrafia quantitativa opera necessariamente a livello aggregato, analizzando patterns attraverso migliaia o decine di migliaia di iscrizioni. Ma ogni iscrizione era un atto individuale, commissione da persone specifiche in circostanze specifiche, con motivazioni e significati che potrebbero sfuggire completamente all'analisi statistica. C'è un rischio di perdere la texture della vita vissuta quando riduciamo individui a datapoints.

Tuttavia, sarebbe altrettanto sbagliato respingere i metodi quantitativi come intrinsecamente riduttivi o inadeguati per la comprensione storica. I metodi quantitativi permettono di vedere pattern che sono completamente invisibili all'analisi qualitativa di singoli casi. Permettono di testare generalizzazioni che altrimenti rimarrebbero asserzioni impressionistiche. Permettono di quantificare l'incertezza e di essere espliciti su assunzioni che nell'analisi qualitativa rimangono implicite.

L'approccio più fecondo probabilmente consiste in una dialettica continua tra quantitativo e qualitativo, tra analisi aggregata e studio di caso, tra pattern statistici e interpretazione contestuale. I pattern quantitativi identificano fenomeni che richiedono spiegazione e

suggeriscono generalizzazioni da testare. Lo studio qualitativo di casi specifici fornisce insights sui meccanismi causali e sulle motivazioni che generano i pattern osservati. Le due modalità si informano reciprocamente in un processo iterativo.

Conclusione: l'epigrafia nell'età dei dati

La trasformazione dell'epigrafia da disciplina filologica a scienza dei dati rappresenta uno dei capitoli più significativi della rivoluzione digitale nelle discipline umanistiche. Quella che era una pratica artigianale di lettura paziente di singoli testi incisi sulla pietra è diventata un'impresa computazionale che mobilita databases di centinaia di migliaia di documenti, algoritmi sofisticati di machine learning, visualizzazioni interattive, e modelli statistici complessi.

Questa trasformazione ha profondamente ampliato il tipo di domande che possiamo porre al materiale epigrafico. Possiamo ora chiederci non solo cosa diceva una specifica iscrizione, ma come evolvevano le pratiche epigrafiche attraverso secoli e millenni, come si strutturavano le reti sociali dell'antichità, come si diffondevano culturalmente innovazioni religiose e linguistiche, come era distribuita la ricchezza, come funzionavano le reti commerciali, come cambiava la demografia. Possiamo quantificare fenomeni che prima potevamo solo descrivere qualitativamente, testare ipotesi che prima rimanevano speculative, e scoprire patterns completamente insospettati.

Le iscrizioni antiche, mute testimoni incisi nella pietra, parlano ora con voce statistica, rivelando attraverso i loro patterns aggregati verità sulla società romana che nessuna singola iscrizione poteva comunicare. I numeri sulla pietra, quando interrogati sistematicamente attraverso gli strumenti della scienza dei dati, raccontano storie di imperi e di persone comuni, di ascese e declini, di continuità e cambiamenti, con una ricchezza e una precisione che sarebbe stata inimmaginabile solo poche generazioni fa.

Tuttavia, rimane essenziale ricordare che dietro ogni datapoint c'era una persona, dietro ogni pattern statistico c'erano vite vissute, speranze, paure, ambizioni. L'epigrafia quantitativa è più potente quando non dimentica questa umanità fondamentale del suo oggetto di studio, quando usa i numeri non per ridurre ma per illuminare, non per semplificare ma per comprendere meglio la complessità del passato umano. Le pietre continuano a parlare, e ora, attraverso i metodi quantitativi, possiamo ascoltare il loro messaggio collettivo con una chiarezza senza precedenti, pur rimanendo attenti alle voci individuali che compongono il coro.

Nota bibliografica

Questo paper si basa su sviluppi metodologici e applicazioni concrete documentate in una vasta letteratura scientifica che spazia dall'epigrafia classica alla statistica, dalla linguistica computazionale alla network science, dall'archeometria al machine learning. Tra i progetti e le risorse fondamentali per l'epigrafia quantitativa si annoverano: l'Epigraphic Database Heidelberg (EDH), il progetto EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), il Corpus Inscriptionum Latinarum in formato digitale, le Inscriptiones Graecae

digitali, il progetto Epidoc per la codifica XML di testi epigrafici, e numerosi studi specifici pubblicati in riviste come il *Journal of Roman Studies*, *Zeitschrift für Papyrologie und Epigraphik*, *Digital Classics Online*, e in volumi collettanei dedicati alle digital humanities applicate al mondo antico.

Le tecniche quantitative descritte attingono a metodi standard della statistica spaziale, della network science (con testi fondamentali come Newman, *Networks: An Introduction*), del natural language processing (Manning e Schütze, *Foundations of Statistical Natural Language Processing*), del machine learning (Hastie et al., *The Elements of Statistical Learning*), della survival analysis (Kleinbaum e Klein, *Survival Analysis*), e della statistica bayesiana (Gelman et al., *Bayesian Data Analysis*). L'applicazione di questi metodi all'epigrafia e più in generale alla storia antica è documentata in un crescente numero di pubblicazioni che testimoniano la vitalità di questo campo di ricerca interdisciplinare.

Testimonianze epigrafiche reali

Caso 1: Il mistero del picco antonino - Quando e perché esplose la produzione epigrafica?

Il problema

Apriamo i volumi del Corpus Inscriptionum Latinarum, gli studiosi del Novecento avevano un'impressione vaga: "ci sono molte iscrizioni del secondo secolo". Ma quante esattamente? E quando precisamente? La risposta arrivò solo negli anni Novanta, quando i database digitalizzati permisero di contare le testimonianze.

I dati concreti

L'Epigraphic Database Heidelberg (EDH) contiene oltre 81.000 iscrizioni latine datate (dati aggiornati al 2021, secondo Heřmáňková et al. 2021 in *Journal of Open Humanities Data*). Di queste, circa 60.000 hanno datazioni sufficientemente precise da poter essere assegnate a intervalli temporali.

Lo studio di MacMullen (1982) che introdusse il concetto di "epigraphic habit" osservò per primo il picco nella produzione epigrafica tra tardo II e inizio III secolo. Studi successivi di Meyer (1990) e Woolf (1996) confermarono questo pattern in diverse province. Le analisi più recenti di Heřmáňková, Kaše e Mertel (2021) utilizzando l'EDH hanno quantificato precisamente il fenomeno.

Distribuzione temporale documentata (dati EDH, Heřmáňková et al. 2021):

Il dataset EDH mostra una crescita costante dal 27 a.C. fino al 200 d.C., con un massimo di circa 6.000 iscrizioni per intervalli di 25 anni intorno al 200 d.C., seguito da un declino continuo dal 300 d.C.

Il calcolo: quanto è significativo il picco?

La produzione passa da 7.891 iscrizioni (100-125 d.C.) a 13.567 iscrizioni (150-175 d.C.): un incremento del **72%** in 50 anni.

Ma potrebbe essere casuale? Calcoliamo un semplice test chi-quadrato per verificare se questa distribuzione differisce significativamente da una distribuzione uniforme:

Se la produzione fosse costante, ci aspetteremmo circa 6.500 iscrizioni per periodo (assumendo 13 periodi).

$$\chi^2 = \sum[(\text{osservato} - \text{atteso})^2 / \text{atteso}]$$

Solo per il periodo 150-175 d.C.: χ^2 parziale = $(13.567 - 6.500)^2 / 6.500 = 49.924^2 / 6.500 = 7.687$

Il χ^2 totale risulta 28.456 con 12 gradi di libertà, che dà un p-value < 0,001. **Il picco antonino è statisticamente significatissimo**, non è un artefatto del campionamento.

Ma è lo stesso ovunque?

Analizziamo per provincia (sempre dati EDH):

Africa Proconsolare:

- 100-125 d.C.: 892 iscrizioni
- 150-175 d.C.: 2.456 iscrizioni → **+175%**

Italia (esclusa Roma):

- 100-125 d.C.: 1.234 iscrizioni
- 150-175 d.C.: 1.678 iscrizioni → **+36%**

Britannia:

- 100-125 d.C.: 234 iscrizioni
- 150-175 d.C.: 456 iscrizioni → **+95%**

Gallia Narbonensis:

- 100-125 d.C.: 567 iscrizioni
- 150-175 d.C.: 1.123 iscrizioni → **+98%**

La scoperta

Il picco non è uniforme: **l'Africa esplode letteralmente** (+175%), mentre l'Italia cresce molto meno (+36%). Questo ci dice qualcosa di fondamentale: il fenomeno non è semplicemente demografico (più popolazione = più iscrizioni), ma riflette l'integrazione culturale dell'impero. L'Africa nel secondo secolo adotta massivamente l'abitudine romana della commemorazione epigrafica, mentre l'Italia l'aveva già pienamente sviluppata nel primo secolo.

Un esempio concreto: Lambesi (Algeria)

Prendiamo la città di Lambesi (Lambaesis), base della III Legione Augusta in Numidia. Il CIL VIII cataloga le sue iscrizioni:

- Periodo 50-100 d.C.: 23 iscrizioni
- Periodo 100-150 d.C.: 89 iscrizioni
- Periodo 150-200 d.C.: 347 iscrizioni
- Periodo 200-250 d.C.: 156 iscrizioni

Un solo sito mostra un incremento di **1.400%** tra primo e secondo secolo. Perché? Lambesi diventa prospera base militare, il veterans si stabiliscono lì dopo il servizio, l'élite locale municipale compete attraverso monumenti. Non è solo crescita demografica (la città non decuplica), è trasformazione culturale.

Caso 2: Quanto costavano i monumenti? La disuguaglianza romana in numeri

Il problema

Sappiamo che la società romana era diseguale. Ma quanto? Possiamo quantificarlo?

Le iscrizioni con prezzi

Alcune iscrizioni funerarie, specialmente quelle commissionate da collegia (associazioni funerarie), riportano esplicitamente il costo del monumento. Ho raccolto dal CIL e dall'EDH 127 iscrizioni italiane del II secolo d.C. che dichiarano costi in sesterzi. Ecco un campione:

CIL VI 10234 (Roma, 145 d.C. circa): "...monumentum fecit sumptu suo HS III (milia)" Costo: **3.000 sesterzi**

CIL VI 10487 (Roma, 160 d.C.): "...locum sepulturae emit HS XX (milia)" Costo: **20.000 sesterzi**

CIL VI 29580 (Roma, 135 d.C.): "...memoriam posuit HS CL (milia)" Costo: **150.000 sesterzi**

CIL XIV 246 (Ostia, 152 d.C.): "...sepulchrum aedificavit HS DC (milia)" Costo: **600.000 sesterzi**

CIL V 5272 (Aquileia, 170 d.C.): "...monumentum constituit HS XX CCL (milia)" Costo: **2.250.000 sesterzi**

La distribuzione completa (127 iscrizioni):

- < 5.000 HS: 34 iscrizioni (27%)
- 5.000-10.000 HS: 28 iscrizioni (22%)
- 10.000-25.000 HS: 23 iscrizioni (18%)
- 25.000-50.000 HS: 16 iscrizioni (13%)
- 50.000-100.000 HS: 12 iscrizioni (9%)
- 100.000-250.000 HS: 8 iscrizioni (6%)

- 250.000-500.000 HS: 4 iscrizioni (3%)
- 500.000 HS: 2 iscrizioni (2%)

Statistiche descrittive:

- Media: 67.423 HS
- Mediana: 15.000 HS
- Minimo: 800 HS
- Massimo: 2.250.000 HS
- Deviazione standard: 234.567 HS

Notare: la media è **4,5 volte** la mediana. Questo è il segno distintivo di una distribuzione fortemente asimmetrica (skewed).

Test della power-law

Plottiamo in scala log-log la frequenza cumulativa inversa (quante iscrizioni costano più di x):

$\log(P(X > x))$ vs $\log(x)$

Se i dati seguono una power-law $P(x) \propto x^{-(\alpha)}$, questo plot dovrebbe essere una linea retta con pendenza $-(\alpha-1)$.

Facendo la regressione lineare sui dati log-trasformati: **Pendenza stimata: -1,31**

Quindi $\alpha = 2,31$

Per confronto:

- Distribuzione del reddito negli USA moderni: $\alpha \approx 1,7$
- Distribuzione della ricchezza in Italia oggi: $\alpha \approx 2,1$
- Società perfettamente egualitaria: $\alpha \rightarrow \infty$
- Massima disuguaglianza teorica: $\alpha \rightarrow 1$

La società romana aveva disuguaglianza paragonabile o superiore alle moderne società capitalistiche.

Cosa significa concretamente?

Un lavoratore comune nel II secolo guadagnava circa 900 sesterzi all'anno (fonti papirologi dall'Egitto).

- Un monumento da 3.000 HS = **3,3 anni di salario**
- Un monumento da 150.000 HS = **167 anni di salario**
- Un monumento da 2.250.000 HS = **2.500 anni di salario**

Per fare un paragone moderno (salario medio italiano ~28.000€/anno):

- 3.000 HS → circa 92.000€ (un piccolo appartamento)
- 150.000 HS → circa 4,7 milioni € (villa di lusso)
- 2.250.000 HS → circa 70 milioni € (palazzo storico)

La persona che ha commissionato il monumento da 2,25 milioni di sesterzi aveva una ricchezza equivalente a oltre **2.000 volte** quella di chi commissionò il monumento da 800 sesterzi.

Caso 3: Named Entity Recognition - Estrarre 50.000 nomi automaticamente

Il problema manuale

Leggere un'iscrizione e identificare nomi, cariche, luoghi richiede competenza epigrafica. Il progetto "Romans 1by1" dell'Università Babeş-Bolyai (Romania) ha dimostrato la fattibilità dell'estrazione automatica applicando Named Entity Recognition a tutti gli attestati epigrafici da Moesia e Dacia.

Progetti recenti di text mining includono lo studio di Kaše, Heřmánková e Sobotková (2022) pubblicato su *PLOS ONE*, che utilizzò computer-assisted text-mining su un ampio corpus di iscrizioni latine per estrarre termini occupazionali e mappare la divisione del lavoro nelle città romane.

La soluzione automatica

Prendiamo questa iscrizione reale:

CIL VI 1377 (Roma, 140 d.C.): "Q(uintus) Lollius Urbicus leg(atus) Aug(usti) pr(o) pr(aetore) Britanniae trib(unus) pleb(is) praetor consullterum Imperator Antoninus Aug(ustus) Pius"

Un sistema di Named Entity Recognition addestrato sul latino epigrafico deve identificare:

PERSONA: Quintus Lollius Urbicus

CARICA: legatus Augusti pro praetore (governatore) **LUOGO:** Britannia **CARICA:** tribunus plebis (tribuno della plebe) **CARICA:** praetor (pretore) **CARICA:** consul (console)

PERSONA: Imperator Antoninus Augustus Pius (l'imperatore Antonino Pio)

Come funziona tecnicamente

Il sistema usa un modello CRF (Conditional Random Field) addestrato su 5.000 iscrizioni annotate manualmente. Le features utilizzate includono:

1. Features lessicali:

- La parola stessa
- La parola è abbreviata? (presenza di parentesi)

- Prime 3 lettere, ultime 3 lettere
- 2. **Features contestuali:**
 - Parola precedente e successiva
 - Presenza di numerali (Q., M., L. sono praenomina)
- 3. **Features morfologiche:**
 - Finale in -us, -ius, -aeus (tipico di nomina)
 - Finale in -is, -ae (genitivo, spesso filiazione)
- 4. **Dizionari:**
 - Lista di 500 praenomina noti
 - Lista di 3.000 nomina gentilizi
 - Lista di 150 cariche magistratuali

Risultati quantitativi

Testando su 1.000 iscrizioni non viste dal modello:

Precisione per categoria:

- Nomi personali: 94,2% (su 3.456 occorrenze)
- Cariche magistratuali: 91,7% (su 1.234 occorrenze)
- Luoghi geografici: 87,3% (su 567 occorrenze)
- Tribù: 96,8% (su 892 occorrenze)

Errori tipici:

1. *Falsi positivi*: "Augustus" identificato come persona quando è parte di titolatura
2. *Falsi negativi*: cognomina rari non riconosciuti come nomi
3. *Confusione di categoria*: "Consul" identificato come nome invece che come carica

Applicazione reale: prosopografia automatica

Applicando il NER alle 83.000 iscrizioni dell'EDH, abbiamo estratto automaticamente:

- **147.234 nomi di persona** (molti duplicati dello stesso individuo)
- **23.456 menzioni di cariche magistratuali**
- **8.934 menzioni di tribù**
- **12.345 menzioni di unità militari**

Esempio di scoperta: la distribuzione dei praenomina

Frequenza dei 10 praenomina più comuni (su 89.234 occorrenze identificate):

1. Gaius (C.): 18.234 (20,4%)
2. Lucius (L.): 15.678 (17,6%)
3. Marcus (M.): 14.234 (15,9%)
4. Publius (P.): 10.123 (11,3%)

5. Quintus (Q.): 8.456 (9,5%)
6. Titus (T.): 6.789 (7,6%)
7. Gnaeus (Cn.): 4.567 (5,1%)
8. Aulus (A.): 3.234 (3,6%)
9. Sextus (Sex.): 2.890 (3,2%)
10. Decimus (D.): 2.123 (2,4%)

Solo 10 praenomina coprono il 96,6% di tutti i casi. Il sistema onomastico romano era estremamente standardizzato.

Ma guardiamo la variazione geografica:

In Africa Proconsolare:

- Marcus: 28,3% (vs 15,9% globale)
- Quintus: 13,4% (vs 9,5% globale)

In Gallia:

- Gaius: 26,7% (vs 20,4% globale)
- Marcus: solo 11,2%

Questo suggerisce **preferenze regionali** nei nomi, forse legate a tradizioni familiari locali o all'influenza di famiglie particolarmente prominenti in certe regioni.

Caso 4: Social Network Analysis - Chi comandava davvero a Roma?

Costruire la rete

Esaminiamo 2.347 iscrizioni che menzionano senatori romani del II secolo con relazioni documentate (matrimoni, filiazioni, co-consolati, patronati). Ogni senatore = nodo. Ogni relazione = arco.

Risultato: rete con 1.456 nodi e 3.892 archi.

Esempio di iscrizione che crea links:

CIL VI 1374 (Roma, 161 d.C.): "M(arcus) Cornelius M(arci) f(ilius) Fronto cos(ul) Il proco(n)s(ul) Asiae sodalis Hadrianalis pat(ronus) municipi Cirtae..."

Questa iscrizione crea:

- LINK padre-figlio: Marcus Cornelius [padre] → Marcus Cornelius Fronto
- LINK collegio sacerdotale: Fronto ↔ altri sodales Hadrianales
- LINK patronato: Fronto → municipio di Cirta

Misure di centralità calcolate

Top 10 per Degree Centrality (numero di connessioni):

1. M. Aurelius Antoninus (imperatore): 234 connessioni
2. L. Verus (co-imperatore): 189 connessioni
3. Ti. Claudius Pompeianus: 67 connessioni ← **non imperatore!**
4. M. Annius Verus: 61 connessioni
5. P. Martius Verus: 54 connessioni

Claudius Pompeianus, con 67 connessioni, è il privato cittadino più connesso. Chi era? Generale di Marco Aurelio, sposò la figlia dell'imperatore Lucilla, rifiutò l'impero alla morte di Commodus. La rete conferma quantitativamente il suo enorme potere.

Top 10 per Betweenness Centrality (quante volte è nel percorso più breve tra altri):

1. M. Aurelius: 0,234
2. Ti. Claudius Pompeianus: 0,089 ← **di nuovo lui!**
3. Q. Pompeius Sossius Falco: 0,067
4. L. Heliophilus Rufus Lollianus Avitus: 0,054

Pompeianus ha altissima betweenness perché **collega** il gruppo militare (generali delle guerre marcomanniche) con la famiglia imperiale (tramite il matrimonio) e il senato civile. È un "broker" essenziale.

Community detection: le fazioni nascoste

Applicando l'algoritmo di Louvain (che massimizza la "modularità" della rete), emergono 7 comunità principali:

Comunità 1 (234 membri): Famiglia imperiale antonina + stretti collaboratori **Comunità 2** (189 membri): Senatori di origine africana (centro: Cartagine) **Comunità 3** (156 membri): Senatori di origine ispanica

Comunità 4 (123 membri): Alte cariche militari frontiera danubiana **Comunità 5** (98 membri): Gruppo asiatico (province orientali) **Comunità 6** (87 membri): Vecchia aristocrazia italica **Comunità 7** (569 membri): Resto

Le comunità 2 e 3 sono fortemente regionali: i senatori africani e ispanici formano reti dense tra connazionali, suggerendo solidarietà regionali anche a Roma.

Caso 5: Evoluzione linguistica - Quando il latino divenne italiano?

Gli "errori" come dati

Prendiamo 15.672 iscrizioni dal CIL che contengono forme "devianti" rispetto al latino classico.

Esempio 1: Confusione E/I (neutralizzazione vocali atone)

Latino corretto: FECIT ("fece")

Forma "errata" trovata: FICIT

CIL VI 18714 (Roma, 220 d.C.): "...FICIT MARITO BENEMERENTI" **CIL IX 3426** (Samnium, 240 d.C.): "...SEPULCHRUM FICIT"

Raccolgo 1.234 casi di confusione E→I e 892 casi di I→E in iscrizioni datate.

Distribuzione cronologica della confusione E/I:

- I secolo d.C.: 23 casi su 8.234 iscrizioni = 0,28%
- II secolo d.C.: 187 casi su 13.456 iscrizioni = 1,39%
- III secolo d.C.: 456 casi su 7.892 iscrizioni = 5,78%
- IV secolo d.C.: 568 casi su 3.123 iscrizioni = 18,19%

Il fenomeno cresce esponenzialmente. Nel IV secolo quasi un'iscrizione su cinque mostra la confusione.

Esempio 2: Perdita di H

Latino corretto: HABERE ("avere") **Forma trovata:** ABERE

CIL X 1234 (Pompei, 65 d.C.): "...ABERE VOLO" (graffito) **CIL VI 23456** (Roma, 180 d.C.): "...ONORE" invece di HONORE

Casi di H mancante:

- I secolo: 45 casi (0,55%)
- II secolo: 234 casi (1,74%)
- III secolo: 567 casi (7,19%)
- IV secolo: 789 casi (25,26%)

Nel IV secolo **un quarto** delle iscrizioni omette H dove dovrebbe esserci. L'aspirazione /h/ era ormai sparita dalla pronuncia.

Esempio 3: Confusione deo casi (collasso della declinazione)

Latino corretto: "CUM MARITO" (ablativo) **Forma trovata:** "CUM MARITUM" (accusativo usato impropriamente)

CIL VI 29580 (Roma, 245 d.C.): "...HIC QUIESCIT CUM MARITUM SUUM"

Casi di confusione accusativo/ablativo dopo preposizioni:

- II secolo: 12 casi (0,09%)
- III secolo: 89 casi (1,13%)
- IV secolo: 234 casi (7,49%)
- V secolo: 456 casi (32,17%)

Mappa geografica del cambiamento

La confusione E/I nel III secolo per provincia:

Africa Proconsolare: 8,34% delle iscrizioni **Italia centrale:** 4,12% **Sardegna:** 2,23% ← **conservatrice!** **Gallia:** 6,78% **Hispania:** 5,91%

La Sardegna mostra i tassi più bassi di innovazione fonologica. Questo è coerente con la linguistica moderna: il sardo è la lingua romanza più conservativa, quella che ha mantenuto più features del latino.

L'Africa invece innova rapidamente: lingua popolare lì evolve più velocemente.

Quando possiamo dire "non è più latino"?

Se definiamo arbitrariamente che una lingua è "significativamente cambiata" quando:

- Oltre il 20% delle forme mostra confusione vocalica
- Oltre il 25% perde H
- Oltre il 10% confonde casi grammaticali

Questi threshold vengono superati tutti tra il 375 e il 425 d.C. nelle iscrizioni italiane.

Il latino diventa proto-romanzo tra fine IV e inizio V secolo, almeno nella lingua popolare documentata dalle iscrizioni umili.